AVERTISSEMENT

LIENS

Code la Propriété Intellectuelle – Articles L. 122-4 et L. 335-1 à L. 335-10
Loi n° 92-597 du 1er juillet 1992, publiée au *Journal Officiel* du 2 juillet 1992

http://www.cfcopies.com/V2/leg/leg-droi.php
http://www.culture.gouv.fr/culture/infos-pratiques/droits/protection.htm

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par Université Toulouse 1 Capitole (UT1 Capitole)

École doctorale : **Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse (EDMITT)**

Présentée et soutenue par

Fériel Boulfani

Le 03 Juin 2021

Caractérisation du comportement des systèmes électriques aéronautiques à partir d'analyses statistiques

---

Spécialité : Mathématiques et Applications

Unité de recherche : UMR 5314 -TSE-R

Directeurs de thèse : **Mme. Anne Ruiz-Gazen et M. Xavier Gendre**

JURY:

M. Hervé Cardot,Professeur des universités, Rapporteur
M. Mathieu Ribatet,Professeur des universités, Rapporteur
Mme Mathilde Mougeot,Professeur des universités, Présidente
M. Jean-Philippe Navarro, Docteur, Examinateur
Mme Anne Ruiz-Gazen,Professeur des universités, Directrice de thèse
M. Xavier Gendre,Professeur des universités, Co-directeur de thèse
Mme. Martina Salvignol, Ingénieur, Invitée

À mon fils Farès.

Remerciements

La thèse de doctorat représente un travail long et laborieux. Mais grâce à la contribution de quelques personnes, ce travail a pu être accompli dans les meilleures conditions. Je tiens par la présente thèse à leur adresser mes plus vifs remerciements.

En premier lieu, j'adresse mes remerciements à mes directeurs de thèse Anne et Xavier, qui ont été toujours d'une grande disponibilité pour m'aider dans mes travaux de recherche. Merci pour votre soutien dans les moments les plus difficiles. Je vous dois beaucoup concernant l'aboutissement de cette thèse.

Je tiens aussi à remercier Hervé Cardot et Mathieu Ribatet d'avoir accepté de rapporter ma thèse et d'avoir contribué à l'amélioration du manuscrit via leurs commentaires constructifs.

Il serait impossible pour moi de ne pas remercier Martina et Philippe, car sans eux cette thèse n'aurait peut-être jamais pu voir le jour. Merci de m'avoir fait confiance en tant que stagiaire dans un premier lieu, puis en tant que doctorante.

Ce travail n'aurait pu être mené à bien sans la disponibilité et l'accueil chaleureux que m'ont témoigné mes collègues du département IYNE et spécialement IYNE1. Je vous présente mes sincères remerciements.

Enfin, je remercie celles et ceux qui me sont chers, ma famille et mes amis, de m'avoir soutenu en dehors des heures de travail tout au long de ces années.

Table des matières

Acronyms

AC Alternating Current
APU Auxiliary Power Unit
CI Confidence Interval
ELA Electrical Load Analysis
EVT Extreme Value Theory
i.i.d. independent and identically distributed
GEV Generalized Extreme Value
GPD Generalized Pareto Distribution
KVA Kilo-Volt-Ampere
min minutes
PP-plot Probability-Probability plot
QQ-plot Quantile-Quantile plot
RAT Ram Air Turbine
sec seconds
UCI Upper Confidence Interval
MSN Material Serial Number
IC Invariant Component
ICA Independent Component Analysis
DTW Dog Teeth Wear
NGCA non-Gaussian component analysis
FOM functional outlier map
DO Directional Outlyingness
PCA Principal Component Analysis
POC Proof of Concept
NN Neural Network
RR Ridge Regression
RF Random-Forest
ReLU Rectified linear unit
GD Gradient Descent
SGD Stochastic Gradient Descent
CART Classification And Regression Tree
FDA Functional Data Analysis
MSE Mean Square Error
OOB Out Of Bag Error
C° degree Celsius
AOG Aircraft On Ground
DT Digital Twin
kts Knots
ft Feet

Introduction

Ce travail de recherche est le résultat d'une collaboration de type Cifre entre le laboratoire TSE-R de l'Université Toulouse 1 Capitole, l'Institut de Mathématiques de Toulouse de l'Université Toulouse 3 et le bureau d'étude du système électrique d'Airbus, dirigé par Madame Martina Salvignol. L'objectif de cette thèse est de répondre à trois problématiques industrielles différentes en utilisant des outils statistiques basés sur les données observées durant les vols. Le présent document est la version publique du manuscrit de thèse qui a été évalué par le jury et qui comportait des parties confidentielles additionnelles.

La principale fonction d'un système électrique est de fournir de l'énergie électrique aux équipements. D'un modèle d'avion à l'autre, l'architecture du système électrique diffère selon le besoin de l'avion (quantité de passagers, type de moteurs, etc.). Pour motiver les travaux présentés dans cette thèse, il est important d'introduire quelques notions sur les systèmes électriques aéronautiques. Nous avons fait une sélection des principales notions utilisées durant cette thèse et qui restent valables sur tous les modèles d'avion.

- **Éléments électriques:** le système électrique se compose de plusieurs éléments essentiels :
  - Sources électriques : elles fournissent de l'énergie électrique aux équipements qui en ont besoin. Dans la Figure 0-1, on illustre un exemple des différentes sources électriques du modèle A320. On compte ici deux générateurs électriques principaux IDG1/IDG2 qui génèrent l'électricité via les moteurs, deux batteries BAT1/BAT2 et les générateurs de secours APUGEN/EMER GEN. Durant le vol les générateurs électriques sont la source principale du système électrique. En complément, des batteries peuvent fournir de l'énergie pendant un temps limité en cas de besoin. Au sol, l'avion peut s'alimenter avec une source externe nommée groupe de parc ;
  - Convertisseurs/transformateurs : ils transforment l'électricité pour obtenir la tension souhaitée et un courant continu ou alternatif ;
  - Éléments de protection : ils permettent de protéger les équipements d'une surconsommation ou d'une surtension ;
  - Éléments de distribution : il s'agit principalement des câbles électriques.

- **Charges électriques:** les charges électriques sont des équipements installés sur des avions et qui consomment de l'énergie électrique pour fonctionner. Les charges sont divisées en deux types, les charges permanentes sont celles qui ont une durée supérieure à 5 minutes et les charges intermittentes sont celles qui ont une durée inférieure à 5 minutes. On parle de charge essentielle si la charge est importante pour terminer le vol, sinon la charge est non essentielle.

- **Mode de fonctionnement:** Le système électrique doit assurer son fonctionnement en mode nominal et dégradé. Le mode nominal signifie qu'on fonctionne sans panne. Quand le système

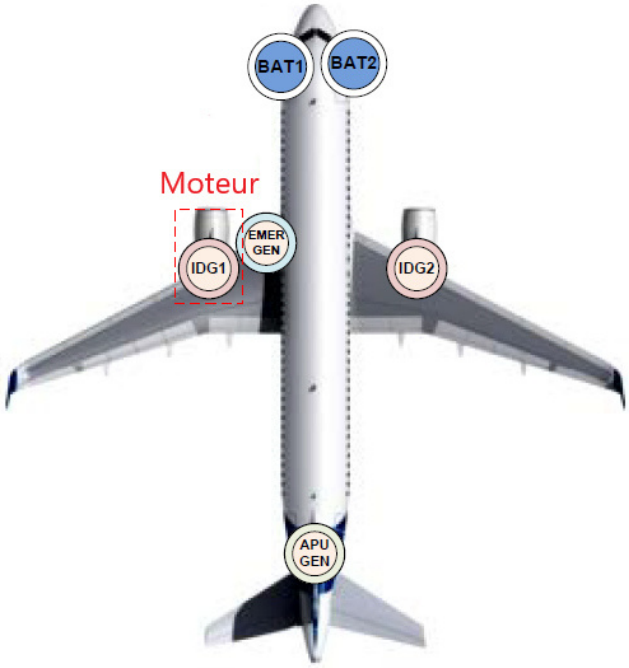

Figure 0-1 – Les différentes sources électriques de la famille A320 (Source Giraud 2014).

a un ou plusieurs composants en panne, on parle de mode dégradé. Durant le mode dégradé, le système continue à fonctionner mais on observe une perte de performance due au délestage (coupure) des charges non essentielles.

- **Phases de vol:** le vol est découpé en plusieurs phases et, selon le modèle d'avion le nombre de phases diffère. Le découpage des phases dépend essentiellement de la vitesse des moteurs, de l'état des générateurs et de l'altitude.

- **Electrical Load Analysis (ELA)**: il s'agit d'un document légal livré avec l'avion à la compagnie aérienne (airline). L'ELA détaille toutes les charges alimentées par le système électrique, par phase de vol, par barre électrique et par générateur. L'ELA estime les charges maximales et opérationnelles par générateur sous les différents modes de fonctionnement. Les charges maximales sont celles où l'utilisation de l'avion est dans des conditions maximales, par exemple, le nombre de passagers est au maximum, tous les fours sont allumés, le dégivrage est allumé, etc. Pour les charges opérationnelles, les conditions d'utilisation suivent les procédures standards de l'avionneur et l'airline. On trouve aussi dans l'ELA les estimations des charges permanentes et intermittentes.

- **Stockage des données:** durant le vol, les données sont enregistrées avec des capteurs et stockées dans des boites d'enregistrements. A l'atterrissage, les données sont transférées sur des plateformes de type cloud et supprimées des boites d'enregistrement. Tout au long de cette thèse, nous avons exploité les données sur deux plateformes, la plateforme Palan-

tir Foundry du groupe Palantir Technologies et la plateforme MAMBA interne à Airbus. Les données ont été prétraitées sur les plateformes puis téléchargées localement pour être exploitées sous Python ou R.

Toutes les problématiques posées par l'industriel concernent des évènements rares mais avec des conséquences lourdes sur les avions. Pour répondre à ces problématiques nous avons appliqué, sur les données d'Airbus, des méthodes statistiques innovantes issues de trois domaines différents. Pour faciliter la lecture, nous avons choisi de traiter chaque problématique séparément. Cette introduction permet de comprendre le contexte industriel et les besoins des ingénieurs avec un résumé qui introduit les différentes approches statistiques utilisées et qui donne les résultats les plus importants. Elle est découpée en trois sections qui correspondent aux trois grandes parties de la thèse. La Section 1 concerne l'estimation de la consommation maximale qu'un générateur doit fournir et s'appuie sur la théorie des valeurs extrêmes. Dans la Section 2, on utilise la théorie des données fonctionnelles pour prédire la température de l'huile dans le but de comprendre le comportement du générateur sous des conditions extrêmes. Pour finir, dans la Section 3, on traite l'anticipation de pannes du générateur électrique en utilisant la méthode "Invariant Coordinate Selection" dans un contexte de données fonctionnelles.

1 Caractérisation de la consommation électrique aéronautique

Au sein des bureaux d'études d'Airbus, l'estimation de la consommation électrique des différentes charges du réseau électrique a toujours été au cœur de toutes les analyses. D'un avion à l'autre, la consommation électrique diffère et dépend essentiellement des charges installées dans l'avion. La précision de l'estimation de la consommation est primordiale car un générateur ne supporte pas une surcharge d'une durée qui dépasse cinq minutes. Au-delà, on perd le générateur, ce qui peut entraîner la surcharge des générateurs de remplacement. De plus, dans certaines phases de vol, on peut aussi perdre le moteur associé au générateur en surcharge, si celui-ci subit un prélèvement élevé quand le moteur est au ralenti (risque de caler le moteur).

De nos jours, l'estimation des consommations électriques par ELA est obtenue en sommant toutes les charges. De la même façon, les générateurs ont été dimensionnés pour fournir la puissance nécessaire pour alimenter toute les charges (somme des charges les plus élevées). Cette méthode repose sur le fait que l'occurrence de ces charges est simultanée, ce qui engendre une estimation conservatrice.

Un premier constat de l'exploitation des données d'avions en service est que la consommation électrique estimée par ELA est toujours supérieure à la consommation réelle. Les ingénieurs du bureau d'études des systèmes électriques se posent des questions sur leur méthode d'estimation

Table 1.1 – Descriptions des groupes. # représente la quantité disponible

Groupe	# avion	# heures de vol	Continent destinations
1	2	10 263	Asie
2	1	1 675	Amérique - Europe
3	4	10 694	Europe
4	2	5 589	Asie
5	5	22 480	Asie
6	2	5 726	Amérique - Europe - Océanie
7	1	1 825	Amérique du nord
8	1	1 793	Europe - Amérique du nord
Total	**18**	**60 045**	-

dans l'ELA et cherchent de nouvelles approches. Un premier travail a été réalisé dans ce sens par Roblot (2012), sur des données d'A380 avec une approche de type simulation des charges électriques. Cette approche n'utilise pas de données réelles de consommation mais une simulation des charges en utilisant la méthode de Monte Carlo et au final donne des résultats loin de la réalité.

Avec la révolution du big data et de l'intelligence artificielle chez Airbus, nous disposons d'une grande quantité de données de vols en service et nous nous intéressons à l'utilisation de ces données dans un cadre statistique pour estimer la consommation électrique maximale. Le but est de revoir le dimensionnement des générateurs, d'améliorer le design des avions opérationnels et de dimensionner les futurs avions électriques.

Cette nouvelle approche repose sur l'utilisation des données en service pour calculer la consommation électrique associée à une probabilité par heure de vol de 10^{-7} puis de la comparer au calcul de l'ELA pour dégager une marge. Le choix de cette probabilité est lié aux procédures de sécurité d'Airbus. Sachant que la probabilité de perdre un générateur est de 10^{-5} par heure de vols, en se positionnant à 10^{-7}, on ne dégrade pas la probabilité de perdre un générateur. Cette marge, si elle est positive, sera intégrée à l'ELA et généralisée à tous les avions opérationnels et aux futurs designs de systèmes électriques.

Nous disposons d'un échantillon de 18 avions de 8 groupes d'avions de type low-cost (bas coûts). Les consommations sont enregistrées chaque seconde sous forme de série temporelle (voir Figure 1-2) pour chaque générateur avec une durée totale de plus de 60 000 heures de vol. Les vols disponibles ont été opérés sur tous les continents du globe et couvrent les quatre saisons météorologiques. Sur la Table 1.1 on reporte la quantité d'avions et le nombre d'heures de vol disponibles par groupe.

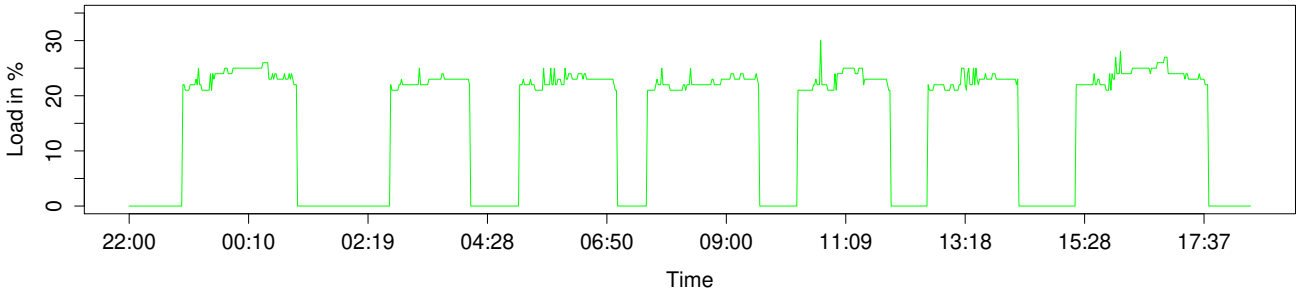

Figure 1-2 – Visualisation de la forme des données disponibles. Consommation électrique en pourcentage de la capacité du générateur pour un avion donné et un générateur donné

1.1 Calcul de quantiles extrêmes

Le but de ce premier travail de recherche est de calculer les quantiles associés à des probabilités d'occurrence d'évènements rares. On définit la variable aléatoire X comme étant la consommation électrique par groupe et F sa fonction de répartition donnée par

$$F(x) = \mathbb{P}(X \leq x),$$

où $\mathbb{P}$ est la mesure de probabilité et x un nombre réel.

Une première approche pour calculer des quantiles est de trouver la fonction de répartition de notre variable aléatoire en utilisant un test d'ajustement à une loi de probabilité donnée. Pour utiliser cette méthode, nous devons faire une approximation empirique de notre fonction de répartition puis, selon sa forme, choisir des lois de probabilité sur lesquelles effectuer des tests d'ajustement. Les distributions empiriques peuvent être utilisées pour sélectionner des lois de probabilité. Sur la Figure 1-3, on donne les distributions empiriques de la consommation électrique de quelques avions. Visuellement, on voit que nos données sont complexes et ne nous permettent pas de sélectionner une loi continue afin d'effectuer les tests d'ajustement. Par ailleurs, même si une distribution est considérée en adéquation avec des données suite au résultat d'un test d'ajustement, elle peut ne pas être satisfaisante pour l'estimation de quantiles extrêmes.

Une deuxième approche qui peut être envisagée en voyant les distributions empiriques est de considérer un mélange de distributions. L'approche de mélange de distributions est utilisée lorsqu'on sait qu'on a une population qui contient deux ou plusieurs sous-populations. Cette méthode est largement utilisée en classification non supervisée en machine learning, et s'avère efficace lorsque le nombre de sous-populations est fini et connu d'avance. Dans notre cas d'application, on voit dans la Figure 1-3 que le nombre de sous-populations diffère d'un avion à un autre. Fixer un nombre de sous-populations différent pour chaque avion n'a pas de sens et ne peut être ex-

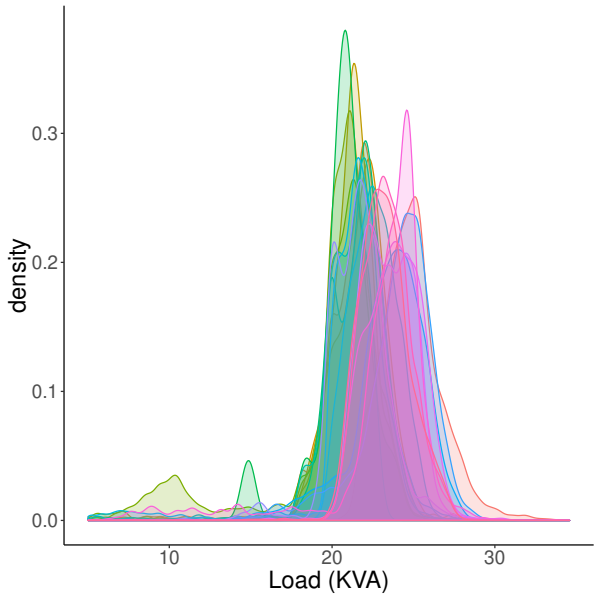

Figure 1-3 – Densité de la consommation électrique du générateur 1 par avions. Chaque couleur représente un avion.

pliqué électriquement puisqu'il s'agit du même modèle d'avion pour toute l'étude. De plus, dans cette approche comme dans la précédente, même si un mélange de lois ajuste bien les données globalement, cet ajustement peut ne pas être satisfaisant pour l'estimation de quantiles extrêmes.

Une approche plus adaptée que les précédentes pour calculer des probabilités d'occurrence d'évènements rares est celle de la théorie des valeurs extrêmes, en anglais Extreme Value Theory (EVT) que nous allons détailler maintenant.

1.2 Théorie des valeurs extrêmes

La théorie des valeurs extrêmes est largement utilisée dans plusieurs domaines scientifiques depuis les 70 dernières années pour modéliser des événements extrêmes ou rares. Les cas d'applications les plus connus sont la prédiction des pics de crues (Morrison and Smith, 2002), des ouragans (Coles et al., 2001), de l'âge maximal que l'être humain pourrait atteindre (Einmahl et al., 2019) ou encore la gestion des risques financiers (Gilli et al., 2006). Cette théorie connaît un grand succès et est appliquée dans des domaines de plus en plus variés. Dans le monde de l'ingénierie, elle est intégrée au moment du design pour augmenter la fiabilité des systèmes. Par exemple, dans le domaine de la corrosion des matériaux, les auteurs Babu and Kondraivendhan (2020) ont utilisé la théorie des valeurs extrêmes pour analyser la corrosion des barres d'armature enrobées dans du béton. Les travaux de Verma et al. (2019) ont utilisé la théorie des valeurs extrêmes pour évaluer le niveau de sécurité structurelle d'un offshore en estimant la distribution extrême de la réponse structurelle sous tous les états opérationnels possibles. Dans le domaine électrique, la théorie des

valeurs extrêmes est utilisée pour estimer des consommations électriques extrêmes comme dans les travaux de Westerlund and Naim (2019) et Ganger et al. (2014).

On retrouve aussi plusieurs cas d'applications dans le domaine de l'aéronautique. Parmi les travaux les plus récents, Larson and Gebre-Egziabher (2017) ont estimé la probabilité d'occurrence d'erreur de mesure dans le système de navigation et Sun et al. (2017) ont utilisé la théorie des valeurs extrêmes pour étudier le spectre de charge pour les pompes hydrauliques des avions.

Pour pouvoir appliquer la théorie des valeurs extrêmes sur nos données, nous devons respecter certaines hypothèses. La première est que les observations de la consommation électrique sont issues de variables aléatoires indépendantes et identiquement distribuées. La deuxième est que chaque variable aléatoire a une fonction de distribution limite non dégénérée. Cette condition est vérifiée pour la majorité des lois de distribution continues (voir Embrechts et al. (2013) pour une liste de lois). Ces deux hypothèses reviennent à supposer a priori que la loi de nos variables aléatoires appartient à une famille de lois paramétriques très générale, ce qui n'est pas une hypothèse forte en pratique.

Dans les ouvrages de De Haan and Ferreira (2006) et Coles et al. (2001), on retrouve toute la théorie des valeurs extrêmes ainsi que des exemples d'applications dans le domaine de l'ingénierie. Ces ouvrages détaillent les deux principales approches de la théorie des valeurs extrêmes, l'estimation des paramètres, le calcul des quantiles extrêmes et le calcul des bornes du support de distribution (points terminaux ou endpoints en anglais) si ces bornes sont finies.

Les deux principales approches de la théorie des valeurs extrêmes sont basées sur la loi d'extremum généralisée et la loi de Pareto généralisée. Ces deux approches peuvent être utilisées pour construire un modèle de valeurs extrêmes de maxima et estimer les paramètres des distributions. Nous proposons un rappel de ces deux approches et de leurs principes.

Loi d'extremum généralisée

La loi d'extremum généralisée, en anglais generalized extreme value distribution (GEV), consiste à découper la série temporelle en blocs de même largeur et à prendre le maximum dans chaque bloc. Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires i.i.d. de fonction de répartition inconnue. On définit

$$M_n = \max\{X_1, \ldots, X_n\},$$

le maximum observé sur un bloc de taille n. En d'autres termes, si n est le nombre d'observations durant 1 heure, alors M_n correspond au maximum observé sur un bloc d'une heure.

Comme indiqué chez Coles et al. (2001), la fonction de répartition asymptotique de M_n est définie par la fonction

$$H(x) = \exp\left\{-(1+\xi\,\frac{x-\mu}{\sigma})^{-1/\xi}\right\}$$

où $1+\xi\,\dfrac{x-\mu}{\sigma} > 0$. Les paramètres $\mu \in \mathbb{R}$ et $\sigma > 0$ correspondent respectivement aux paramètres de position et de dispersion. Le troisième paramètre $\xi \in \mathbb{R}$ correspond au paramètre de forme appelé aussi indice des valeurs extrêmes. Selon la valeur de ξ, la fonction H va converger vers la distribution de Fréchet, Gumbel ou Weibull.

Loi généralisée de Pareto

La loi généralisée de Pareto, en anglais generalized Pareto distribution (GPD) consiste à choisir un seuil assez élevé et à sélectionner les excès qui dépassent ce seuil.

Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires i.i.d. de fonction de répartition inconnue. On définit un seuil u et les excès $X_i - u$, pour $i \in \{1, \ldots, n\}$.

Pour μ, $\sigma > 0$ et ξ donnés et pour un seuil u suffisamment grand, la fonction de répartition de $X - u$ conditionnelle en $X > u$ peut être approchée par

$$H(x) = \begin{cases} 1 - \left(1 + \xi\dfrac{x}{\beta}\right)^{-1/\xi} & \text{si } \xi \neq 0, \\ 1 - \exp\left(-\dfrac{x}{\beta}\right) & \text{si } \xi = 0, \end{cases}$$

avec $x > 0$ et $\beta = \sigma + \xi(u - \mu) > 0$ le paramètre de dispersion modifié.

Les paramètres, les quantiles extrêmes, les points terminaux de F ainsi que leurs intervalles de confiance sont estimés via une approche asymptotique où le seuil u est remplacé par une suite de statistiques d'ordres supérieurs qui dépendent de n (voir De Haan and Ferreira (2006) pour plus de détails). Dans la pratique, pour pouvoir utiliser ces résultats asymptotiques, il faut que le nombre d'observations n soit grand mais aussi que le ratio n_u/n où n_u est le nombre d'observations qui dépassent le seuil u, soit petit (voir Einmahl et al. (2019) pour une application détaillée). Le choix de u nécessite un arbitrage entre le biais et la variance. Généralement, u est sélectionné en utilisant des outils graphiques de diagnostique.

1.3 Application de la théorie des valeurs extrêmes sur les consommations électriques aéronautiques

A notre connaissance, l'utilisation de la théorie des valeurs extrêmes sur la consommation du système électrique dans le domaine de l'aéronautique est nouvelle. En utilisant les résultats présentés dans De Haan and Ferreira (2006), nous pouvons estimer des quantiles extrêmes et des points

terminaux de façon ponctuelle ou par intervalles de confiance.

Pour estimer la consommation électrique d'un générateur donné, nous avons choisi de grouper les avions selon leurs configurations. Nous avons aussi choisi de séparer les consommations selon si elles sont permanentes ou intermittentes et selon le mode de fonctionnement du générateur, nominal ou dégradé.

Nous avons choisi la GPD et la méthode POT (Peaks Over Threshold) de Pickands III et al. (1975) pour estimer nos quantiles. La méthode GEV n'a pas été considérée car nous avons rencontré des difficultés pour découper les vols en blocs et ajuster les modèles avec la GEV. Pour des raisons historiques l'approche POT est généralement privilégiée et il existe plusieurs outils graphiques pour faciliter le choix du seuil.

En utilisant les résultats de De Haan and Ferreira (2006) nous pouvons estimer des quantiles extrêmes, des points terminaux et aussi construire des intervalles de confiance pour ces derniers.

Pour estimer la consommation électrique d'un générateur donné, nous avons choisi de grouper les avions selon leurs configurations. Nous avons aussi choisi de séparer les consommations selon si elles sont permanentes ou intermittentes et selon le mode de fonctionnement du générateur, nominal ou dégradé.

Comme le but final est de pouvoir généraliser les résultats aux avions non observés, et cela indépendamment de leur configuration, nous avons choisi d'estimer le ratio de la consommation électrique par rapport à la valeur maximale théorique donnée dans l'ELA. En divisant la consommation électrique observée par la valeur théorique, nous estimons la proportion d'électricité consommée par l'avion comparée à sa valeur théorique. En effet, nous avons constaté qu'en utilisant la méthode GPD directement sur des données de consommations électriques, nous estimons la consommation de l'avion le plus chargé. Or, plus l'avion est chargé, plus il va consommer et plus ses maxima seront sélectionnés par l'ajustement de la GPD. Le calcul des ratios permet de pallier ce problème. Si un ratio est inférieur à 1, cela signifie que l'ELA surestime la consommation réelle et qu'il sera éventuellement possible de dégager de la marge correspondant à l'écart entre la valeur 1 et le ratio estimé.

L'application de la théorie des valeurs extrêmes sur ces ratios (dénotés par Y_i, $i \in \{1, \ldots, n\}$) nous permettra d'estimer des quantiles extrêmes (ratios extrêmes) et des points terminaux (ratios terminaux) de ces ratios ainsi que des intervalles de confiance indépendamment de la configuration de l'avion.

Pour pouvoir appliquer la théorie des valeurs extrêmes sur la consommation électrique permanente nous devons distinguer les consommations permanentes des intermittentes. A partir des enregistrements de consommation totale par générateur à la seconde dont nous disposons, l'identification est impossible. Pour pouvoir extraire les consommations permanentes, nous avons choisi de calculer des moyennes pour lisser les intermittents sur une fenêtre de temps de longueur

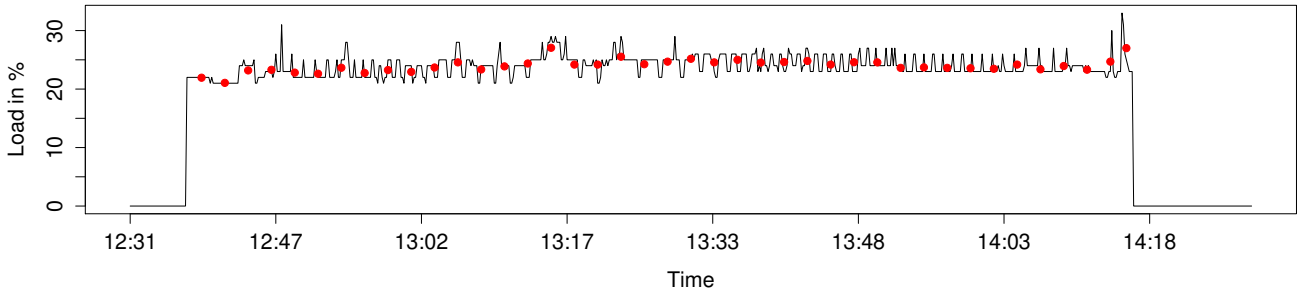

Figure 1-4 – Calcul de la moyenne de la consommation électrique pour un générateur donné sur une fenêtre de temps donné pour un vol donné.

$T = 150$ secondes :

$$X_k = \frac{1}{T}\sum_{t=1}^{T} Y_{(k-1)T+t}, \quad k \in \{1, \dots, \tau\}$$

où $\tau = \lfloor n/T \rfloor$ et $\lfloor \cdot \rfloor$ représente la partie entière. Avec cette transformation, on récupère de nouvelles observations issues de variables aléatoires i.i.d. X_k.

Sur la Figure 1-4, on visualise le lissage d'un vol donné. Les points rouges seront divisés par l'ELA théorique et considérés comme étant les réalisations de notre variable aléatoire X qui représente "le ratio de la consommation électrique permanente par la consommation électrique théorique".

Pour estimer les paramètres de la GPD nous avons utilisé le package *extRemes* (voir Gilleland et al. (2016)) sous le logiciel R avec la méthode du maximum de vraisemblance. Tous les résultats ont donné des ξ négatifs, ce qui signifie que le support de la loi des maxima a une borne supérieure et que les points terminaux à droite des ratios calculés existent bien.

Sur ces points terminaux et quantiles extrêmes, un intervalle de confiance est construit avec un niveau de risque de 10^{-3}. Dans la Table 1.2, nous présentons les résultats liés aux points terminaux dénotés par $\hat{X}^*$ et leurs intervalles de confiance dénotés par $\mathrm{CI}_{10^{-3}}$ par groupe et par phase. On observe une petite variation entre les groupes et les phases. Le plus grand ratio observé correspond à 83%, ce qui correspond à une marge de 17 %.

Pour pouvoir généraliser ce résultat à tous les groupes d'avions, nous avons utilisé le test d'égalité des points terminaux développé par Einmahl et al. (2019). Ce test a été appliqué aux points terminaux des groupes dans le but de voir si les points terminaux des groupes sont égaux. Nous avons appliqué ce test sur la phase de vol et sur la phase au sol séparément. Le résultat des tests a conclu à ne pas rejeter l'hypothèse que les points terminaux sont égaux avec un risque de première espèce de 5% pour les deux phases. Le non-rejet de cette égalité démontre que la plus

Table 1.2 – Points terminaux et bornes supérieures de leur intervalle de confiance par groupe et par phase pour un générateur donné.

Groupe	Phase vol		Phase sol	
	$\hat{X}^*$	$\text{CI}_{10^{-3}}$	$\hat{X}^*$	$\text{CI}_{10^{-3}}$
1	68.5	75.9	69.8	75.6
2	70.3	72.1	68.7	73.4
3	69.7	71.8	74.4	82.6
4	66.6	76.1	70.6	77.8
5	69.6	72.5	72.7	76.6
6	71.4	75.5	72.4	76.1
7	67.5	70.3	70.3	72.2
8	72.2	77.5	66.2	72.8

grande valeur que peut prendre le ratio ne dépend pas des groupes d'avions considérés.

Suite à ce résultat, nous avons décidé de mélanger tous les groupes et d'estimer les points terminaux par phase. Nous avons trouvé un point terminal de 75% pour la phase vol et de 80% sur la phase sol. On constate encore une fois que les points terminaux sont proches et cela nous conduit à tester l'égalité des points terminaux entre les deux phases. Le test d'égalité n'est pas rejeté avec un risque de première espèce de 5%. Grâce à ces résultats, nous concluons qu'il est possible de considérer que les résultats obtenus sur les ratios ne dépendent ni des groupes d'avions ni des phases.

Finalement, un point terminal global a été estimé en mélangeant tous les groupes et sans distinction entre les phases de vol et de sol. Sur la Figure 1-5, on donne la borne supérieure de l'intervalle de confiance du point terminal global avec des niveaux de risque allant de 0.05 à 10^{-12}. La ligne pointillée représente la plus grande valeur observée pour l'intervalle de confiance (86%) et la ligne hachurée le ratio égal à 100%. On constate, comme on pouvait s'y attendre, que plus le niveau de risque est faible, plus le ratio est grand et risque de dépasser 100%. Pour un niveau de risque à 10^{-3}, on a un point terminal global de 80% ce qui est en cohérence avec les résultats précédents et confirme une marge de 20% sur l'ELA dans le mode nominal et pour les charges permanentes.

En utilisant une approche statistique, nous avons pu chiffrer une surestimation que les ingénieurs suspectaient mais n'arrivaient pas à estimer. Cependant, ce résultat ne prend en compte que les consommations permanentes et le mode nominal et ne peut donc pas être utilisé pour dimensionner les générateurs. On peut toutefois se servir des résultats pour alléger certaines limitations ou améliorer le design du réseau électrique. Pour compléter l'évaluation du réseau électrique, nous devons aussi intégrer des avions qui ne sont pas low-cost et surtout étendre l'étude aux charges intermittentes et au mode dégradé qui sont des cas dimensionnant pour un générateur électrique. Les travaux de cette étude ont été soumis au journal "CEAS aeronautical journal".

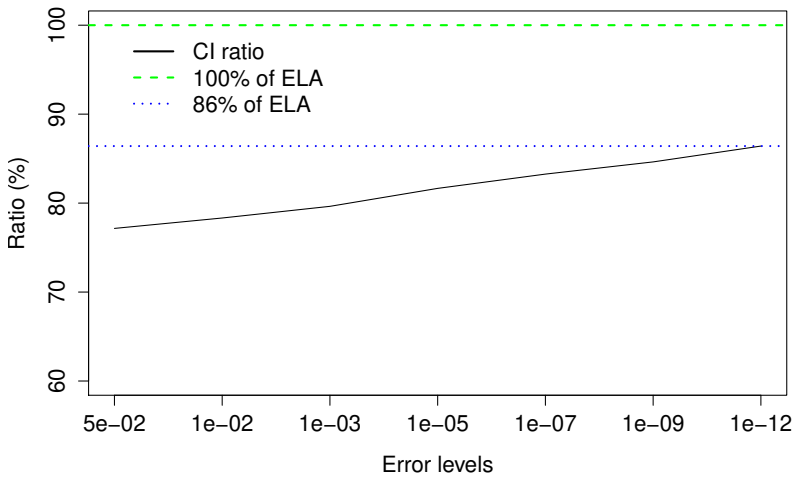

Figure 1-5 – Variation de l'intervalle de confiance du point terminal global en fonction du niveau de risque α.

Notons aussi qu'un outil a été développé sur Palantir Foundry avec le support de l'équipe d'Airbus India. Cet outil opérationnel a été conçu pour permettre aux ingénieurs d'appliquer facilement les méthodes développées dans cette thèse sur des données de vols disponibles chez Airbus.

2 Prédiction de la température de l'huile d'un générateur électrique aéronautique

Le système électrique aéronautique est le système le plus complexe des systèmes aéronautiques car il s'agit d'une ressource partagée et en interaction permanente avec les autres systèmes. De plus, on observe une augmentation de la demande d'électrification de certaines fonctionnalités dans les avions. Chaque modification du système électrique nécessite une reconception (redesign) du système avec des simulations basées sur des équations physiques et/ou avec des simulations sur avion. L'étude de redesign permet d'être sûr que le générateur ne sera pas dégradé après les modifications du système électrique.

L'huile du générateur permet de refroidir le système électrique et sa température peut être utilisée pour mesurer le bon fonctionnement du générateur. Une température élevée ou basse reflète un mauvais fonctionnement du générateur. Pour comprendre l'interaction du générateur avec les autres systèmes, la prédiction de la température de l'huile sous des conditions extrêmes (par exemple, très haute ou très basse altitude, température extérieure extrême, etc.) peut nous aider à valider les modèles électriques et anticiper les défauts de design avant la production.

Nous disposons de données enregistrées durant les vols avec une fréquence d'un enregistrement

par seconde. Nous avons choisi de prédire la température de l'huile à un temps donné en fonction de l'historique des variables explicatives. Les modèles non fonctionnels ont donné des résultats de moindre qualité comparés aux modèles fonctionnels, ce qui explique notre choix.

Dans cette partie de la thèse, nous détaillons l'utilisation des outils d'analyse de données fonctionnelles développés par Ramsay and Silverman (2005) pour prédire une sortie scalaire. Nous détaillons aussi les procédures de prédiction les plus utilisées et leur régularisation avec la technique du dropout. Une application de cette approche a été développée pour anticiper les pannes de générateurs électriques.

2.1 Prédiction dans un contexte fonctionnel

La prédiction d'une sortie scalaire dans un contexte fonctionnel nécessite une représentation des données dans l'espace des fonctions de carré intégrable.

On suppose que les données observées sont des vecteurs de fonctions appartenant à l'espace de Hilbert $L^2 = L^2([0,1], dt)$ où dt désigne la mesure de Lebesgue sur $[0,1]$. Dans ce qui suit, on considère le produit scalaire usuel,

$$\forall f, g \in L^2, \langle f, g \rangle_{L^2} = \int_0^1 f(t)\, g(t)\, dt.$$

Soit $\{\xi_d\}_{d\in\mathbb{N}}$ une base orthonormale de L^2. La projection d'une fonction $f \in L^2$ sur cette base orthonormale est donnée par

$$\forall t \in [0,1], f(t) = \sum_{d\in\mathbb{N}} c_d\, \xi_d(t),$$

où

$$c_d = \langle f, \xi_d \rangle_{L^2} = \int_0^1 f(t)\, \xi_d(t)\, dt.$$

Dans notre application, les fonctions explicatives sont observées de façon discrète, donc les coefficients c_d sont estimés sur une grille régulière de pas $\frac{1}{T}$ sur $[0,1]$ par

$$\hat{c}_d = \frac{1}{T} \sum_{i=1}^{T} f\left(\frac{i}{T}\right)\, \xi_d\left(\frac{i}{T}\right).$$

Pour réduire la dimension, on applique une troncature en sélectionnant les $D \in \mathbb{N}$ premiers coefficients de chaque variable explicative. Sur les coefficients sélectionnés, on applique des procédures habituelles de régression. Plusieurs bases peuvent être considérées (voir Ramsay and Silverman (2005)). Dans notre application, on choisit les bases courantes qui sont de Fourier et de Haar.

Soit $C \in \mathbb{R}^{nq}$ la matrice des coefficients résultant de la troncature, où n représente le nombre d'observations et q le nombre total de coefficients retenus et $y \in \mathbb{R}$ le scalaire a prédire. Nous cherchons un prédicteur f_ω basé sur les entrées C où ω est un vecteur de paramètres à estimer. La qualité d'un prédicteur est évaluée par le risque

$$L(f_\omega) = \mathbb{E}_B \left[\ell(y, f_\omega(\mathrm{c})) \right],$$

avec $\mathrm{c} \in \mathbb{R}^q$ et B est la distribution de probabilité jointe de y et C. Dans notre application, nous utilisons la fonction de perte quadratique $\ell(y, y') = (y - y')^2$ pour évaluer f_ω. Le risque empirique de l'estimateur est défini par

$$\widehat{L}(f_\omega) = \frac{1}{n} \sum_{k=1}^{n} \left(y_k - f_\omega(\mathrm{c}_k) \right)^2 .$$

Le but est de trouver le meilleur algorithme qui minimise l'équation

$$\min_{\omega \in \Omega} \widehat{L}(f_\omega). \tag{2.1}$$

Nous avons sélectionné les approches par régression linéaire, réseau de neurones et forêts aléatoires. Ces algorithmes peuvent avoir une bonne prédiction sur les données initiales (données d'entraînement) mais sont de mauvaise qualité pour prédire de nouvelles données (données de test). Ce phénomène est appelé overfitting (surapprentissage), pour faire face à cette situation il est nécessaire d'avoir recours à des techniques de régularisation.

Nous avons choisi la méthode du dropout comme technique de régularisation. Cette méthode a été popularisée par Srivastava et al. (2014) et est très utilisée en pratique dans les applications des réseaux de neurones. Elle consiste à mettre quelques poids du réseaux de neurones à zéro de manière aléatoire avec une probabilité $p \in [0, 1)$, ce qui peut être vu comme une régularisation par le bruit. L'application du dropout à l'étape d'entraînement peut être généralisée à tous les algorithmes en mettant quelques entrées à zéro avec une probabilité p. Nous nous sommes intéressés aux travaux de Arora et al. (2020), Mianjy et al. (2018) et Neyshabur et al. (2015) pour comprendre l'effet du dropout sur le risque de chaque estimateur qu'on propose.

Ci-dessous, nous donnons un bref récapitulatif des algorithmes sélectionnés et leur version régularisée par la technique du dropout :

- La **régression linéaire** suppose une relation linéaire entre les variables explicatives et la variable à prédire. Le modèle est défini par $f_\omega(\mathrm{c}) = \mathrm{w}^\top \mathrm{c} + w_0$ où $\omega = (\mathrm{w}, w_0) \in \mathbb{R}^q \times \mathbb{R}$. L'équation (2.1) devient

$$\min_{(\mathrm{w}, w_0) \in \mathbb{R}^q \times \mathbb{R}} \left\{ \frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \mathrm{c}_k - w_0 \right)^2 \right\}.$$

Le dropout pour la régression linéaire met à zéro les variables d'entrées avec la probabilité p. La version du risque empirique sous l'espérance du dropout de ce modèle est donnée par

$$\widehat{L}_{\text{drop}}(f_\omega) = \widehat{L}(f_\omega) + \frac{p}{1-p}\|\Gamma \mathrm{w}\|^2 \quad \text{avec} \quad \Gamma^2 = \text{ddiag}\left(\frac{1}{n}\sum_{k=1}^{n} \mathrm{c}_k \mathrm{c}_k^\top\right),$$

où ddiag est une matrice diagonale dans $\mathbb{R}^{d\times d}$.

On voit que le dropout régularise en ajoutant le terme $p(1-p)^{-1}\|\Gamma \mathrm{w}\|^2$ à $\widehat{L}(f_\omega)$ similaire à une régularisation de Tikhonov à la différence que la matrice Γ dépend des variables d'entrées. Pour des données normalisées on retrouve une pénalité quadratique de type ridge. Ce type de lien est connu et a été discuté dans les travaux de Bishop (1995), Wang and Manning (2013), Heinze et al. (2014) et chez Wager et al. (2013) avec une approche dropout semblable à notre approche.

- Le **réseau de neurones** est une collection de neurones connectés entre eux pour prédire y. Chaque connexion a un poids et chaque neurone a une fonction d'activation qui définit l'état du neurone. Un perceptron est un réseau de neurones qui contient plusieurs couches (layers) et chaque couche contient des neurones dont les sorties sont les entrées de la couche suivante. Plus le nombre de couches est grand, plus le nombre de paramètres à estimer est grand et plus on risque d'être confronté au phénomène de surapprentissage.

  Pour un perceptron à une couche, on dispose d'une couche d'entrée (input layer), une couche cachée (hidden layer) et une couche de sortie (output layer). Le prédicteur f_ω est obtenu par la combinaison linéaire des poids d'entrée $(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)})$ et des poids de sortie (w, w_0) du perceptron,

$$\forall \mathrm{c} \in \mathbb{R}^q,\ f_\omega(\mathrm{c}) = \mathrm{w}^\top \varphi\left(\mathrm{V}^{(1)}\mathrm{c} + \mathrm{v}_0^{(1)}\right) + w_0,$$

  où $\omega = (\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}, \mathrm{w}, w_0) \in \mathbb{R}^{q_1\times q} \times \mathbb{R}^{q_1} \times \mathbb{R}^{q_1} \times \mathbb{R}$ est la collection de paramètres à estimer et φ la fonction d'activation appliquée aux neurones de la couche cachée.

  L'équation (2.1) pour un perceptron à une couche devient

$$\min_{(\mathrm{V}^{(1)},\mathrm{v}_0^{(1)},\mathrm{w},w_0)\in\mathbb{R}^{q_1\times q}\times\mathbb{R}^{q_1}\times\mathbb{R}^{q_1}\times\mathbb{R}} \left\{\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \mathrm{w}^\top\varphi\left(\mathrm{V}^{(1)}\mathrm{c}_k + \mathrm{v}_0^{(1)}\right) - w_0\right)^2\right\},$$

  et le risque empirique sous l'espérance du dropout est donné par

$$\widehat{L}_{\text{drop}}(f_\omega) = \widehat{L}(f_\omega) + \frac{p}{1-p}\left\|\Gamma\left(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}\right)\mathrm{w}\right\|^2$$

  avec

$$\Gamma^2\left(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}\right) = \text{ddiag}\left(\frac{1}{n}\sum_{k=1}^{n}\varphi\left(\mathrm{V}^{(1)}\mathrm{c}_k + \mathrm{v}_0^{(1)}\right)\varphi\left(\mathrm{V}^{(1)}\mathrm{c}_k + \mathrm{v}_0^{(1)}\right)^\top\right).$$

On voit que la régularisation dropout sur un perceptron à une couche agit comme une régularisation de Tikhonov sur les sorties de la couche cachée.

Pour un perceptron à plusieurs couches, le dropout peut être appliqué à chaque couche cachée avec des probabilités de dropout différentes. Une écriture explicite du terme que le dropout ajoute à $\widehat{L}(f_\omega)$ reste compliquée mais nos résultats expérimentaux sur des données de vols montrent que la probabilité de dropout depend du nombre de couches et du nombre de neurones dans les couches cachées. Plus le modèle a de neurones plus il demande une grande probabilité de dropout.

- Le **CART** est un arbre de décision construit par divisions successives de l'espace des données en deux parties distinctes afin de prédire y par de simples moyennes. Cette étape est répétée jusqu'à la satisfaction d'un critère d'arrêt (stopping rule). La décision de diviser ou le choix du point de division sont sélectionnés automatiquement par l'algorithme de sorte à minimiser la fonction de perte. Le critère d'arrêt définit à quel moment on arrête de diviser les données. Plusieurs règles existent comme le nombre d'individus dans une feuille de l'arbre ou un critère de variance. L'élagage (pruning) proposé par Breiman et al. (1984), permet de supprimer des branches peu représentatives pour avoir un sous-arbre meilleur en terme d'erreur de prédiction.

  Les forêts aléatoires sont une méthode d'agrégation d'arbres de type CART. Pendant l'entraînement, à chaque division, m variables tirées aléatoirement sont utilisées pour construire ces arbres. Les forêts aléatoires peuvent être vues comme la version régularisée de CART par la technique du dropout. Nos expérimentations sur les données de vols ont montré qu'on peut appliquer une probabilité de dropout allant jusqu'à 0.4 avec une petite perte dans la performance de la procédure.

2.2 Détection de pannes de générateurs

Dans cette application, nous avons exploité l'idée que la température de l'huile reflète le bon fonctionnement du générateur pour prédire un fonctionnement anormal. Le principe est d'entraîner les algorithmes avec des vols qui ne contiennent pas d'anomalies et de les tester sur des vols qui contiennent des pannes. Le but ainsi recherché est d'avoir une grande erreur de prédiction sur les vols qui précèdent la panne.

Pour cela, nous avons sélectionné 600 vols de 6 avions. Pour chaque vol, nous avons sélectionné les variables explicatives données dans la Table 2.1. On veut prédire la température de l'huile y à un instant donné en fonction de l'historique de longueur T des variables explicatives. Chaque vol a été divisé en segments de longueur T pour construire une matrice X où chaque ligne représente un segment de vol.

Table 2.1 – Variables explicatives utilisées pour prédire la température de l'huile du générateur

Variable	**Déscription**	**Unité**
y	Température de l'huile du générateur	C°
X^1	Vitesse du moteur	Knot (kts)
X^2	Température statique de l'air	C°
X^3	Température totale de l'air	C°
X^4	Vitesse de l'air calculée	kts
X^5	Altitude	ft
X^6	Charge électrique du générateur	KVA

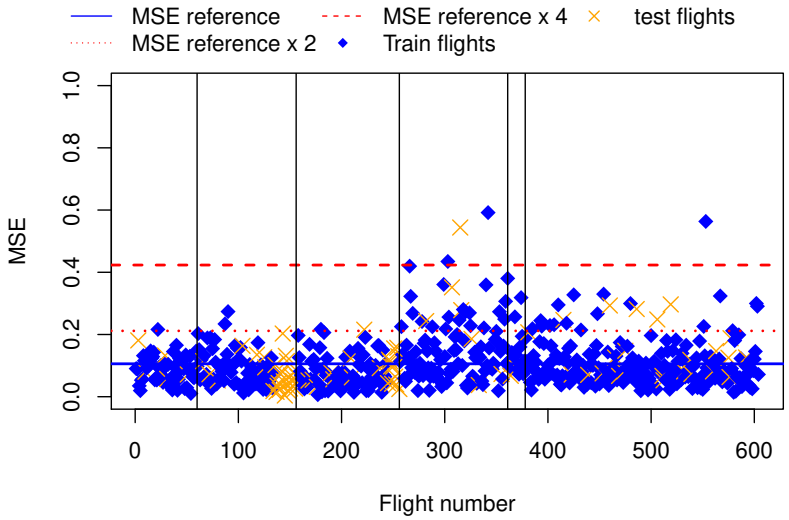

Figure 2-6 – Erreur de prédiction par vol en utilisant un réseau de neurones et la base de Fourier.

La matrice des segments a été projetée sur une base de Fourier et de Haar dans lesquelles on a appliqué des troncatures en ne gardant que les D premiers coefficients. Toutes les troncatures résultantes ont été injectées dans les algorithmes de machine learning : modèle linéaire, forêts aléatoires et réseaux de neurones. Le meilleur résultat observé est pour les coefficients de Fourier avec un réseau de neurones.

Dans la Figure 2-6, on donne en ordonnée l'erreur quadratique moyenne (MSE) par vol et, en abscisse, les indices des vols ordonnés par date. Chaque point représente un vol, les points avec la forme de diamant bleu sont des vols d'entraînement et les points en forme de croix orange sont des vols de test. On a ajouté dans cette figure une ligne bleue qui représente la moyenne des erreurs de tous les vols (MSE de référence) et on a ajouté des repères de distances en rouge, à 2 et 4 fois le MSE référence. On voit que ce modèle a presque toutes ses erreurs inférieures au repère de distances égal à 4 fois le MSE de référence et qu'elles sont presque toutes concentrées autour du MSE de référence.

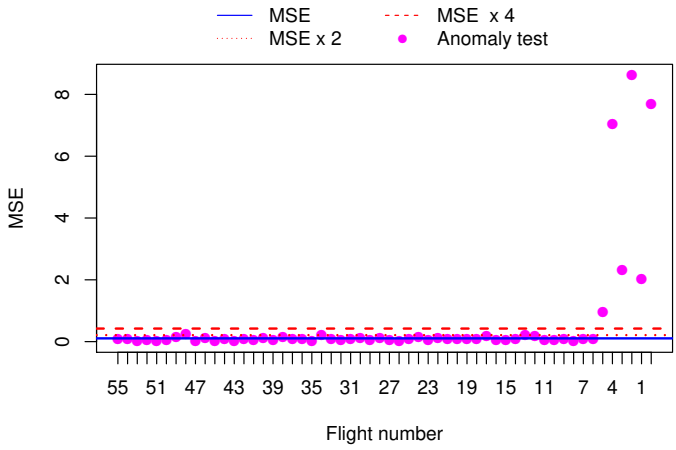

Figure 2-7 – MSE par vol pour détecter la perte du générateur du cas 1

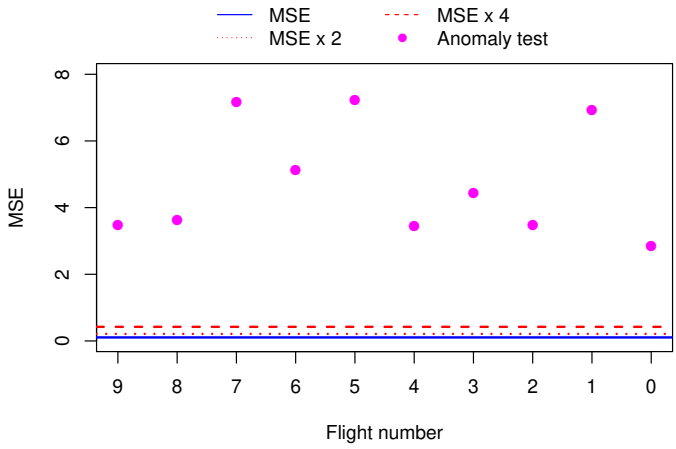

Figure 2-8 – MSE par vols pour détecter la perte du générateur du cas 2

Deux cas de perte de générateur ont été testés sur ce modèle en injectant les vols qui précèdent la panne dans notre algorithme pour calculer l'erreur de test. Les résultats de ces cas sont donnés dans les Figures 2-7 et 2-7 où on retrouve les mêmes repères de distances que sur la Figure 2-6. Les vols sont numérotés à l'aide d'un décompte jusqu'à la panne, ce qui fait que le vol numéro 0 représente la perte du générateur.

Pour le cas 1, on voit clairement sur la Figure 2-7 que les vols qui précèdent la perte ont une erreur plus grande que des autres vols. Les anomalies dans le fonctionnement du générateur sont visibles 5 vols avant sa perte. Ce délai permet d'enclencher la maintenance sans perturber le planning de vols de l'avion. Dans le cas 2, nous ne disposons que de 10 vols qui précèdent la panne. Sur la Figure 2-7, on voit que les 9 vols ont une erreur très grande et annonce la perte du générateur.

Ce travail a été présenté et publié au congrès international GC-ElecEng (Global Congress on Electrical Engineering) en septembre 2020. L'algorithme développé a été présenté comme un digital twin (jumeau digital) pour anticiper les pannes du générateur électrique.

3 Détection d'anomalies dans le fonctionnement d'un générateur électrique aéronautique

Le système électrique aéronautique, comme tout équipement, subit des dégradations qui annoncent l'arrivée d'une défaillance du système. De toutes les défaillances que le système électrique peut avoir, la panne du générateur électrique est celle qui engendre le plus de difficultés. Les pannes du générateur électrique sont rares mais nécessitent beaucoup de temps pour y remédier, ce qui entraîne des retards ou l'annulation de vols. Le plus souvent le générateur est remplacé. Le coût d'un nouveau générateur s'ajoute donc aux pénalités de retard, ce qui fait de cette panne la plus coûteuse.

Les ingénieurs du bureau d'étude du système électrique s'intéressent à la possibilité d'anticiper les pannes du générateur pour pouvoir réduire les délais et les coûts. En utilisant une sélection de données enregistrées durant le vol et une approche statistique, on souhaite pouvoir alerter l'airline d'une panne imminente pour qu'elle procède à une maintenance de l'avion.

La plupart des approches statistiques liées à la détection d'anomalies ne sont pas adaptées aux cas de fichiers de données avec une petite proportion de données atypiques (inférieure à 5%). Les travaux d'Archimbaud et al. (2018) sur la méthode Invariant Coordinate Selection (ICS) démontrent que cette méthode est efficace pour détecter des données atypiques dans un contexte de faible proportion d'anomalies. Les auteurs prouvent aussi que la distance de Mahalanobis fonctionne souvent mal et proposent d'utiliser la méthode ICS car elle permet de sélectionner des composantes pertinentes à la détection d'observations atypiques. Cette méthode est similaire à l'Analyse en Composante Principale (ACP) car elle permet de faire une réduction de dimension en sélectionnant un petit nombre de composantes. Mais au lieu d'interpréter les valeurs propres en termes de variance comme pour l'ACP, il faut les interpréter en termes de kurtosis (coefficient d'aplatissement d'une distribution). Contrairement à l'ACP, l'ICS a plus de chance de garder la structure d'atypicité des données et elle est affine invariante alors que l'ACP est juste orthogonalement invariante.

Dans notre cas d'application, nous avons choisi la représentation en données fonctionnelles parce qu'on observe pour chaque vol des enregistrements en fonction du temps à un pas fin. La représentation fonctionnelle des données est très souvent utilisée de nos jours, on en trouve des applications dans des domaines très différents comme la météorologie (voir Beyaztas and Yaseen, 2019 et Suhaila et al., 2011), la medecine (voir Gruen et al., 2017 et Ratcliffe et al., 2002) et le

contrôle de qualité (voir Torres et al., 2020 et Millán-Roures et al., 2018).

La détection d'anomalies sur des données fonctionnelles multivariées a fait l'objet de nombreux travaux récents (Rousseeuw et al. (2018), Staerman et al. (2019), Dai et al. (2020) et Lejeune et al. (2020)). Mais, le plus souvent, les exemples étudiés concernent au maximum 3 variables fonctionnelles (Kuhnt and Rehage (2016), Rousseeuw et al. (2018), Dai and Genton (2019), Dai et al. (2020), Staerman et al. (2019) et Lejeune et al. (2020)) et utilisent une approche de profondeur fonctionnelle (functional depth) ou pseudo-profondeur fonctionnelle qui coûte cher en calcul pour des vols de longue durée (voir Hubert et al. (2015), Kuhnt and Rehage (2016) et Dai et al. (2020)).

Dans le domaine de l'aéronautique, les applications de la détection d'anomalies sur des données de vols traitent le plus souvent l'aspect multivarié sous forme de série temporelle (voir Li et al. (2015) et Li et al. (2016), Memarzadeh et al. (2020)). A notre connaissance, peu de travaux utilisent l'approche fonctionnelle pour la détection d'anomalies dans le domaine de l'aéronautique. On peut citer Jarry et al. (2020) qui utilisent l'ACP fonctionnelle en univarié pour détecter des individus atypiques en utilisant des enregistrements de radar, et Lejeune et al. (2020) qui utilisent une approche d'extraction de forme des atypiques pour détecter des pannes moteur.

L'application d'ICS à des données fonctionnelles n'a encore jamais été étudiée dans un cadre de données fonctionnelles. Récemment, Li et al. (2019) et Virta et al. (2020) ont proposé une généralisation sur des données fonctionnelle mais dans un contexte d'analyse en composantes indépendantes (ICA). La généralisation d'ICS sur des données fonctionnelles peut se faire de différentes manières. Nous proposons deux approches d'ICS fonctionnelle. La première, dite point-wise ICS, consiste à faire un ICS standard à chaque temps puis à calculer des indicateurs agrégés à partir des distances. La deuxième approche, dénommée global ICS, consiste à faire une projection des variables fonctionnelles sur une base orthonormale et à appliquer une troncature pour ne garder que quelques coefficients par fonction. Sur ces coefficients, on applique une seule fois ICS standard.

Dans cette partie du manuscrit de thèse, on applique les méthodes d'ICS généralisées à des données fonctionnelles à deux jeux de données pour lesquels les observations atypiques sont connues. Le premier cas concerne la détection d'anomalies pour anticiper des pannes d'un générateur électrique aéronautique dans un contexte de maintenance prédictive. Le deuxième cas concerne la détection de semiconducteurs défectueux dans un contexte de contrôle de qualité. Dans cette introduction, nous avons choisi de ne présenter que le cas aeronautique.

3.1 Détection d'anomalies avec ICS standard

Soit un jeu de données $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ avec p variables et n observations, où $'$ représente la transposée. Une matrice de dispersion est une matrice $p \times p$, symétrique, définie positive et équivariante par transformation affine. Une matrice $\mathbf{V}(\mathbf{X}_n)$ est équivariante par transformation

affine si et seulement si

$$\mathbf{V}(\mathbf{X}_n\mathbf{A} + \mathbf{1}_n\mathbf{b}') = \mathbf{A}'\mathbf{V}(\mathbf{X}_n)\mathbf{A},$$

où $\mathbf{A}$ est une matrice $p \times p$ de rang plein, $\mathbf{b}$ un vecteur de taille p et $\mathbf{1}_n$ un vecteur de taille n rempli de 1.

Dans la littérature, il existe plusieurs matrices de dispersion (voir Nordhausen and Tyler (2015)). Pour ce qui nous concerne, nous considérons la paire de matrices constituée de la matrice de variance-covariance empirique classique et de la matrice basée sur les moments d'ordre quatre. La matrice variance-covariance est définie par

$$\mathrm{COV}(\mathbf{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

où $\bar{\mathbf{x}}$ représente la moyenne empirique. La matrice de moment d'ordre 4 est définie par

$$\mathrm{COV}_4(\mathbf{X}_n) = \frac{1}{(p+2)n}\sum_{i=1}^{n} r_i^2(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

où $r_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathrm{COV}(\mathbf{X}_n)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$ est le carré de la distance de Mahalanobis.

ICS consiste à diagonaliser simultanément la paire de matrices $\mathbf{V}_1(\mathbf{X}_n)$ et $\mathbf{V}_2(\mathbf{X}_n)$. Cette diagonalisation conduit à une matrice $\mathbf{B}(\mathbf{X}_n)$ de taille $p \times p$ et une matrice diagonale $\mathbf{D}(\mathbf{X}_n)$ telles que :

$$\mathbf{B}(\mathbf{X}_n)\mathbf{V}_1(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{I}_p \quad \text{et} \quad \mathbf{B}(\mathbf{X}_n)\mathbf{V}_2(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{D}(\mathbf{X}_n)$$

où $\mathbf{I}_p$ est la matrice identité de taille $p \times p$. La matrice $\mathbf{D}(\mathbf{X}_n)$ contient les valeurs propres de $\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)$ dans un ordre décroissant, tandis que les lignes de la matrice $\mathbf{B}(\mathbf{X}_n) = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ contiennent les vecteurs propres tels que :

$$\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{B}(\mathbf{X}_n)'\mathbf{D}(\mathbf{X}_n).$$

En utilisant n'importe quel estimateur de position équivariant par transformation affine, $\mathbf{m}(\mathbf{X}_n)$, les scores sont calculés par

$$\mathbf{Z}_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)' = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}(\mathbf{X}_n)')\mathbf{B}(\mathbf{X}_n)'$$

où les $\mathbf{Z}_n$ correspondent aux composantes affine invariantes.

Comme démontré par Archimbaud et al. (2018), la norme Euclidienne $\sqrt{\mathbf{z}_i'\mathbf{z}_i}$ est égale à la distance de Mahalanobis pour l'observation i, $i = 1, \ldots, n$. Mais la distance de Mahalanobis n'offre pas la possibilité de réduction de dimension.

Dans le cas d'une faible proportion d'atypiques, les propriétés théoriques d'ICS nous aident à

nous concentrer sur les composantes invariantes associées aux plus grandes valeurs propres (voir Archimbaud et al. (2018) pour plus de détails).

Une fois que le nombre de composantes invariantes k est sélectionné, la dernière étape à faire est d'identifier les observations atypiques. Pour cela, pour chaque observation $i = 1, \ldots, n$, on calcule sa distance au carré appelée distance ICS, à partir des k premières composantes

$$(\text{ICS distance})^2_{i,k} = \sum_{j=1}^{k} \left(z_i^j\right)^2 \tag{3.1}$$

où z_i^j est la $j^{\text{ème}}$ coordonnée du score $\mathbf{z}_i$.

Une observation est étiquetée comme atypique si sa distance ICS calculée sur les k premières composantes dépasse un certain seuil (cutoff).

3.2 Détection d'anomalies sur des données fonctionnelles multivariées

Pour le global et le point-wise ICS, on suppose que les données sont des vecteurs de fonctions de carré intégrable sur $[0, 1]$ par rapport à la mesure de Lebesgue dt. On note l'espace de Hilbert de ces fonctions par $L^2 = L^2([0, 1], dt)$. Dans ce qui suit, on considère le produit scalaire,

$$\forall f, g \in L^2, \langle f, g \rangle_{L^2} = \int_0^1 f(t)\, g(t)\, dt.$$

On dispose de jeux de données avec p variables. Chaque variable a une dimension de $n \times T$ où n est le nombre de fonctions observées et T est le nombre de points temporels observés. En pratique, T peut être différent d'une observation à une autre ou d'une variable a une autre. Pour remédier à ce problème, il faut passer par des techniques d'alignement temporel.

Global ICS

La première étape de global ICS est de faire une projection de nos fonctions sur une base orthonormale. La projection d'une fonction $f \in L^2$ sur la base orthonormale $\{\xi_d\}_{d\in\mathbb{N}}$ est définie par

$$\forall\, t \in [0, 1],\ f(t) = \sum_{d\in\mathbb{N}} c_d\, \xi_d(t), \text{ avec } c_d = \langle f, \xi_d \rangle_{L^2}.$$

L'estimation de c_d dans $[0, 1]$ avec un pas de $1/T$ est donnée par

$$\hat{c}_d = \frac{1}{T} \sum_{t=1}^{T} f\left(\frac{t}{T}\right) \xi_d\left(\frac{t}{T}\right).$$

Parmi les nombreuses bases possibles (Ramsay and Silverman, 2005), on considère les bases de Fourier et les B-splines qui sont couramment utilisées. Contrairement à la base de Fourier, la base des B-splines n'est pas orthonormale et l'estimation de ses coefficients nécessite une adaptation en utilisant la matrice de Gram. Dans notre application d'ICS, on a utilisé l'estimation des moindres carrés sans prendre en compte la matrice de Gram (voir Subsection 5.1 de Virta et al. (2020) pour plus de détails sur l'implémentation de ICS avec la matrice de Gram). Pour réduire la dimension, une troncature est appliquée pour ne garder que les premiers $D \in \mathbb{N}$ coefficients de chaque variable. De cette troncature découle un jeu de données de dimension $n \times pD$ au lieu de $n \times pT$.

Le choix de D est basé sur le nombre d'observations n et, par expérience, nous utilisons une règle de base qui contraint D a être plus petit que $n/(10p)$. On applique l'ICS standard sur le jeu de données obtenu.

Pour la sélection du nombre de composantes invariantes, k, nous avons utilisé l'éboulis des valeurs propres comme proposé par Archimbaud et al. (2018) en cherchant un coude dans la décroissance des valeurs propres. Cette méthode est simple et permet de sélectionner peu de composantes.

Les observations sont étiquetées atypiques si elles dépassent le cutoff proposé par Archimbaud et al. (2018). Ce cutoff est basé sur des simulations de Monte Carlo de valeurs propres calculées à partir de données suivant une distribution Gaussienne. Ces simulations génèrent plusieurs échantillons pour lesquels on calcule les distances ICS au carré. Le cutoff représente le quantile $1 - \gamma$ des distances ICS au carré. Dans notre cas d'application, on a utilisé la valeur par défaut fixée à $\gamma = 2.5\%$.

Point-wise ICS

Pour Point-wise ICS, on applique ICS standard sur les fonctions alignées en chaque point de temps $t = 1, \ldots, T$. Cela nous conduit à p composantes par temps et, au final, à pT composantes. À chaque temps t, on doit sélectionner $k(t)$ composantes. Pour automatiser la procédure, nous avons utilisé le test proposé par Nordhausen et al. (2017) qui conduit à choisir $k(t)$ composantes avec un niveau de signification de 1%.

Pour chaque observation i, on calcule ensuite $(\text{ICS distance})^2_{i,k(t)}(t)$ en utilisant l'expression (3.1) pour chaque t. Pour résumer l'information et étiqueter les observations atypiques, on propose d'adapter l'approche Functional Outlier Map (FOM) de Rousseeuw et al. (2018). Pour chaque

observation, on calcule la moyenne (fICS) et la variabilité (vICS) des distances ICS au carré :

$$\text{fICS}_i = \frac{1}{T}\sum_{t=1}^{T} (\text{ICS distance})^2_{i,k(t)}\,(t)$$

$$\text{vICS}_i = \frac{\text{stdev}_i}{1 + \text{fICS}_i}$$

où stdev_i représente l'écart-type de $(\text{ICS distance})^2_{i,k(t)}\,(t)$.

Dans nos applications, nous avons constaté que les composantes sélectionnées $k(t)$ peuvent varier beaucoup d'un temps à l'autre. Pour pallier ce problème, nous avons choisi de diviser le carré des distances ICS par le nombre de composantes sélectionnées par temps dans le calcul de fICS et vICS.

Pour visualiser l'information et la résumer, on utilise le FOM, qui est un outil graphique sur lequel on représente fICS et vICS. Une grande valeur de fICS correspond a une fonction avec un comportement atypique qui dure dans le temps. Une grande valeur de vICS correspond a une fonction avec un comportement atypique sur quelque segments du temps.

Pour étiqueter les observations atypiques, nous avons utilisé le cutoff décrit par Rousseeuw et al. (2018) avec un quantile adapté à chaque cas d'application.

3.3 Cas d'applications

Dans les applications considérées, nous comparons les résultats des deux approches d'ICS fonctionnelle. Nous discutons le choix du nombre de composantes à sélectionner, le cutoff et, en plus pour global ICS, le choix du nombre de coefficients à garder.

Pour chaque application nous connaissons les observations qui sont atypiques, ce qui va nous permettre d'évaluer les deux approches. Sachant que la quantité des observations atypiques est très faible, une bonne méthode sera une méthode qui détecte toutes les observations atypiques (vrais positifs) avec peu de fausses détections (faux positifs).

Cas d'application d'ICS sur la maintenance prédictive

Cette application concerne le suivi durant le vol du comportement d'un générateur électrique d'un avion donné. Le but est de lancer une alerte avant que le générateur ne tombe en panne.

On dispose de $n = 590$ vols avec des durées différentes. Le vol numéro 591 représente le vol où on perd le générateur de l'avion. Les quatre vols qui précèdent la perte du générateur sont considérés comme des vols anormaux. Pour chaque vol, 6 variables ($p = 6$) ont été identifiées pouvant expliquer le comportement du générateur durant les vols. Dans la Table 3.1, on retrouve

Table 3.1 – Variables sélectionnées pour détecter des anomalies dans les données de vols

Déscription	**Unité**
Température de l'huile du générateur	C°
Vitesse du moteur	Knot (kts)
Température statique de l'air	C°
Vitesse de l'air calculée	kts
Altitude	ft
Charge du générateur	KVA

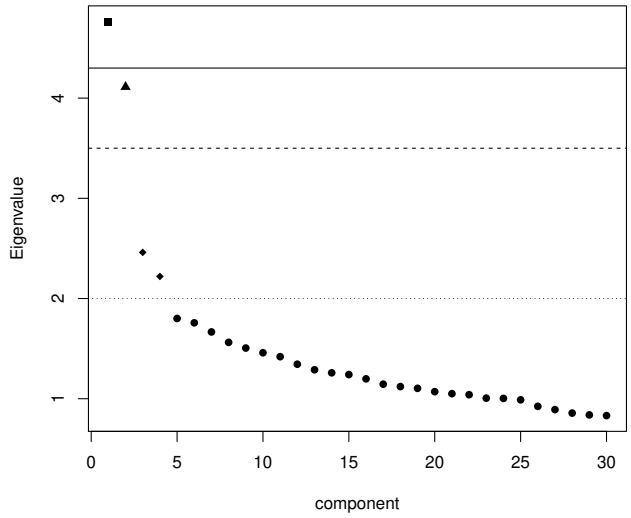

Figure 3-9 – **Jeu de données Aeronautique** - Eboulis des valeurs propres pour global ICS.

les paramètres détaillés. Chaque vol a été aligné par phase de vol puis reconstruit. Dans cette application, les vols ont été alignés sur une durée $T = 2900$ secondes.

Pour global ICS, les vols alignés ont été projetés sur un base de Fourier. Les 5 premiers coefficients ($D = 5$) ont été gardés pour chaque variable ce qui donne un ensemble de données de dimension 590×30, ce qui satisfait la règle de $D < n/(10p)$. Les résultats sont similaires avec la base des B-splines, mais on observe une différence dans les résultats si on fait varier D.

Pour le choix de k, nous avons utilisé l'éboulis des valeurs propres de la Figure 3-9 où on identifie 3 valeurs possibles pour k, $k = 1$ (forme carrée), $k = 2$ (triangle) and $k = 4$ (diamant). Pour chaque valeur de k, on a calculé les distances ICS au carré et représenté ces distances sur la Figure 3-10. Les vols sont ordonnés par date et la ligne verticale hachurée représente la perte du générateur. Les vols détectés comme anormaux sont associées aux mêmes formes que celles utilisées pour l'éboulis des valeurs propres (carré quand le vol est détecté avec la première composante, triangle avec les

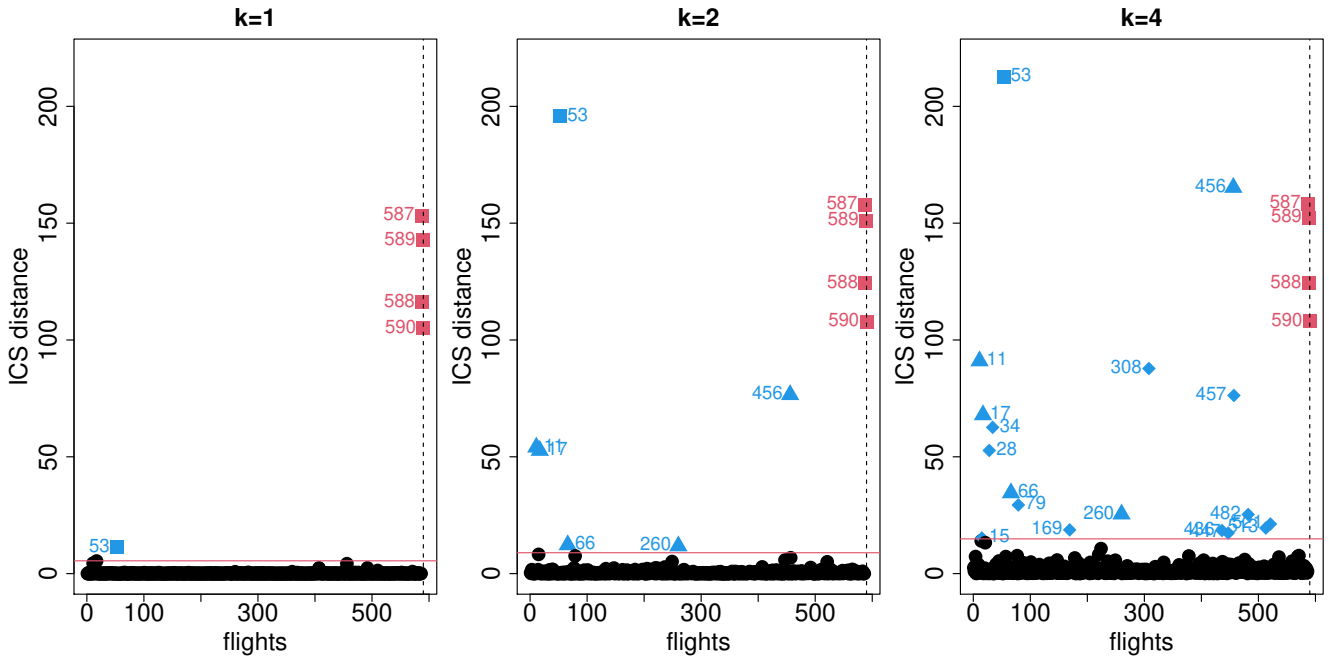

Figure 3-10 – **Aéronautique data set** - Global ICS distances avec $k = 1$ (resp. $k = 2$ et $k = 4$) composantes sélectionnées sur la gauche (resp. au centre et à droite) du panel et un cutoff associé à un quantile de 0.975.

deux premières composantes et diamant avec les 4 premières composantes). Les vrais positifs sont colorés en rouge alors que les faux positifs sont en bleu. Les cutoffs pour chaque groupe sont représentés par une ligne horizontale rouge.

La valeur $k = 4$ détecte une quantité de vols supérieure à la limite de 2% d'observations atypiques dans le data set, ce qui fait qu'on ne considère que les valeurs $k = 1$ et $k = 2$. Avec la valeur $k = 1$, on détecte les 4 vols anormaux sans aucun faux positif. Le caractère atypique des ces vols s'explique par le comportement anormal de la variable vitesse du moteur avant la perte du générateur, comme détaillé dans la troisième partie de la thèse.

Le résultat de l'application de point-wise ICS sur les vols alignés est donnée à la Figure 3-11 où les vrais positifs sont colorés en rouge et les faux positifs sont en bleu. Pour chaque temps, ICS standard a été appliqué avec une sélection automatique du nombre k de composantes. Dans le graphique de gauche, on retrouve le graphique appelé FOM avec un cutoff calculé pour un quantile d'ordre 0.999995. Le graphique du milieu représente un indicateur qui prend la valeur 1 si le vol est détecté atypique et 0 sinon. Tandis que le graphique de droite donne les valeurs de fICS par vol.

En utilisant cette méthode, on voit qu'on détecte tous les vols anormaux mais avec six faux positifs. On remarque que parmi ces 6 faux positifs, 4 vols sont aussi détectés par global ICS avec $k = 2$ (voir Figure 3-10).

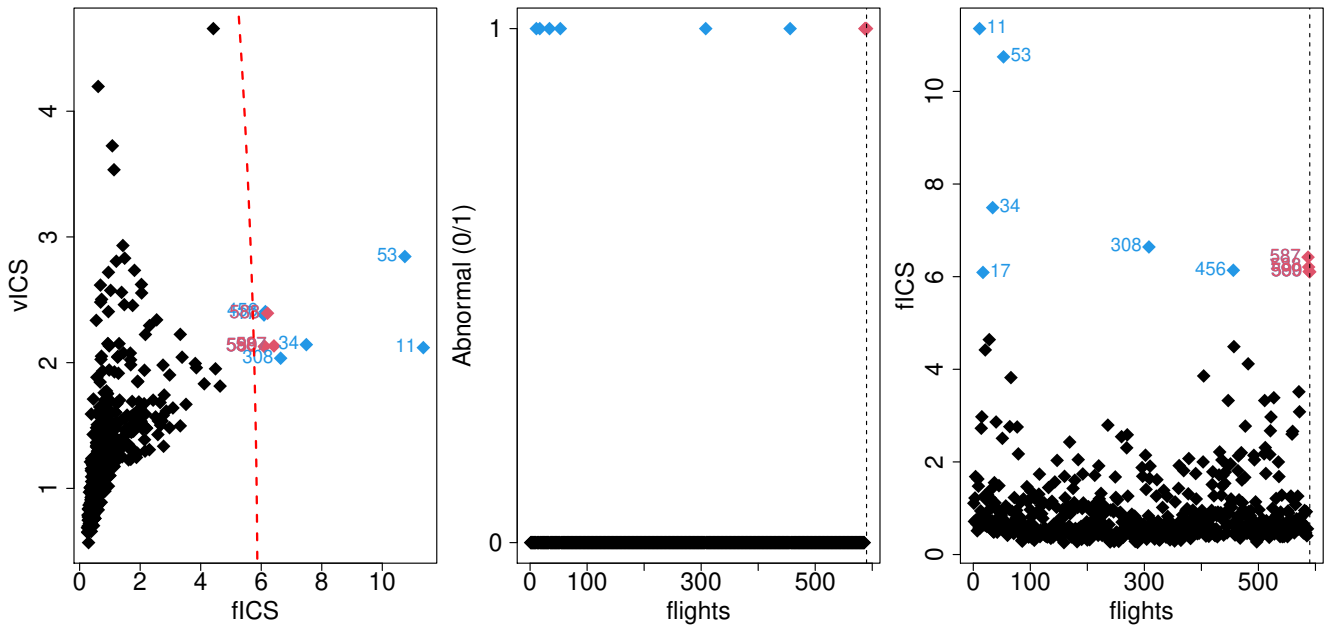

Figure 3-11 – **Jeu de données Aéronautique** - Point-wise ICS pondéré. Gauche : FOM avec un cutoff calculé avec un quantile d'ordre 0.999995, Milieu : identification des anomalies, Droite : fICS par vol.

Global et point-wise ICS n'ont pas donné exactement les mêmes résultats mais ont détecté plusieurs vols atypiques en commun. Pour global ICS, nous avons observé que, pour $D > 11$, les vols anormaux ne sont plus détectés.

Ce travail a été appliqué sur d'autres cas de pannes de générateur et a donné d'aussi bons résultats que pour cette application. L'algorithme développé pour détecter ces pannes est en cours de déploiement pour permettre à des ingénieurs qui assurent le suivi des générateurs de l'utiliser.

4 Organisation du manuscrit

Dans cette introduction, nous avons détaillé le contexte industriel, les problématiques et introduit les approches statistiques utilisées. Chacune des trois parties de cette thèse contient deux chapitres dont un est rédigé sous forme d'article et où nous détaillons les approches et leurs applications. Les parties sont indépendantes et peuvent être lues séparément.

La partie 1 est écrite sous forme d'article consacré à l'estimation de la consommation maximale qu'un générateur doit fournir. Cet article a été soumis à la revue CEAS (Council of European Aerospace Societies).

La partie deux traite la prédiction de la température de l'huile d'un générateur électrique dans un contexte de données fonctionnelles. Le premier chapitre détaille la théorie des données fonction-

nelles et la régularisation des procédures de prédiction. Le second chapitre est sous forme d'article et consiste en une application du premier chapitre sur des données de vols. Cet article a été publié et présenté au congrès international GC-ElecEng (Global Congress on Electrical Engineering) en septembre 2020.

La dernière partie est dédiée à la détection d'anomalies dans un contexte de données fonctionnelles en utilisant la méthode "Invariant Coordinate Selection" (ICS). Cette partie est écrite sous forme d'article consacré à l'adaptation de l'application d'ICS à des données fonctionnelles. Cet article a été soumis au journal "Econometrics & Statistics".

Part I

Aeronautical electrical consumption characterization

Sommaire

Introduction

The electrical system for aircraft represents a collection of electrical components powered by a generator and the generator itself is supplied by the engine.

An electrical overload that is greater than five minutes may result in the loss of the generator and under some conditions the loss of the engine. Because of that the estimation of consumption of the electrical system should be very accurate.

Actually the electrical consumption is estimated and reported in the Electrical Load Analysis (ELA) document. Each aircraft has it ELA and it is provided to the airline at the time of aircraft delivery. In the ELA the maximal consumption is given for the permanent and intermittent loads in the nominal and degraded mode.

ELA estimates maximal consumption by summing the highest loads in the most unfavourable conditions. In this Part, we use a statistical approach to estimate the maximal consumption based on operational measurements and challenge the ELA approach. The idea is to prove that the ELA overestimates the maximal consumption as the loads do not trigger at the same time. Using the Airbus safety procedure we define the maximal consumption as the quantile associated to the probability 10^{-7}.

To this end, we use 18 operational aircraft from the same family. From these aircraft, we recover 60000 flight hours of electrical consumption by generator. On this consumption we apply the extreme value theory to estimate a maximal consumption for permanent load and build a confidence interval around this consumption.

The procedure and the preprocessing of the data is detailed for one generator in Chapter 1 which is reprint of the paper *A statistical approach for sizing an aircraft electrical generator using extreme value theory* submitted to CEAS aeronautical journal.

The comparison between the estimation of the maximal consumption using extreme value theory and the ELA for the permanent load gives a positive gap of 20% which means that the ELA overestimates the maximal consumption for the permanent load.

Chapter 1

Extreme value theory to estimate maximal electrical consumption of generator

This chapter represents an article submitted to CEAS aeronautical journal. This journal is devoted to publishing results and findings in all areas of aeronautics-related science and technology as well as reports on new developments in design and manufacturing of aircraft, rotorcraft, and unmanned aerial vehicles. Due to confidentiality issues, the airlines, the aircraft model and the generator that we use in this work are anonymised.

A statistical approach for sizing an aircraft electrical generator using extreme value theory

Fériel Boulfani, Xavier Gendre, Anne Ruiz-Gazen and Martina Salvignol.

Abstract

The sizing of aircraft electrical generators mainly depends on the electrical loads installed in the aircraft. Currently, the generator capacity is estimated by summing the critical loads, but this method tends to overestimate the generator capacity. A new method to challenge this approach is to use the electrical consumption recorded during flights and study the distribution of operational ratios between the actual consumption and the theoretical maximum consumption then size the future aircraft generators by applying a ratio to the theoretical value. This paper focuses on the application of extreme value theory on these operational ratios to estimate the maximal capacity utilization of a generator. A real data example is provided to illustrate the approach and estimate

extreme quantiles and the right endpoint of the distribution of the ratios together with their approximate confidence interval in the nominal configuration. In all situations the right endpoint is proven to be finite and does not depend on the use procedures.

Keywords

Electrical load analysis; Aeronautic electrical system; Generalized Pareto distribution; Quantile estimation; Endpoint estimation; Diagnostics for threshold selection.

1.1 Introduction

Driven by the demand to reduce emissions, the aviation industry pushes toward the concept of more electrical aircraft and, ultimately, an all-electrical aircraft Seresinhe and Lawson (2015). Thus, the electrical network will be more in demand. A new network should be designed, and a new electrical-intensive architecture implemented.

The Electrical Load Analysis (ELA) is an airworthiness requirement. For a given aircraft, it describes the electrical network and shows the total theoretical electrical consumption by generators for the different flight phases and different operational modes. In the ELA, the electrical consumption is computed by summing the component loads under the most unfavourable conditions to get the maximal consumption and under normal operating conditions to get the operational consumption.

The ELA is provided to the airline at the time of aircraft delivery. The airline must use this report to evaluate the effects of equipment changes on the electrical network to avoid electrical overload.

To avoid oversizing the future electrical network, aircraft manufacturer has to assess the current network and re-evaluate the needs based on operational measurements. According to several recent operational measurements of the electrical network, the theoretical power consumption appears to be overestimated as illustrated in Figure 1.1-1. This figure shows the proportion of electrical consumption with respect to electrical capacities over time for one generator during a given flight. A large difference is observed between the theoretical maximum consumption given by ELA and the real consumption.

Using operational measurements, we want to justify that the maximal observed consumption is smaller than the maximal consumption given by ELA. The main reason is that the electrical loads do not operate all at the same time whereas they are considered simultaneously in the ELA.

A preliminary work has been done by Roblot (2012) using Monte Carlo algorithms to simulate

Figure 1.1-1 – Example of the consumption in percentage of the capacity of generators as a function of time for a given flight

the electrical load behavior. This approach is based on simulations and differs from ours as our objective is focused on the extreme behavior of the observed electrical consumption. The approach developed hereafter is based on the Extreme Value Theory (EVT). EVT provides statistical tools to estimate extreme quantiles and right endpoints under two hypotheses. First, the observations are considered as independent and identically distributed (i.i.d.) realizations of random variables. Second, the probability distribution belongs to the domain of attraction of some extreme value distribution. Under these hypotheses, we derive extreme quantiles and endpoints together with their confidence interval. Note that extreme quantiles (resp. endpoints) are values such that the probability of getting a larger value is extremely small (resp. equal to zero).

The distribution assumption is not restrictive and can be checked for many well known distributions including the uniform on interval and the normal ones (see Embrechts et al. (2013)). The results are asymptotic in the sense that they are valid for large sample size. The parametric extreme value distribution is obtained by looking at the limit distribution of standardized maxima. This result is comparable to the central limit theorem that considers the asymptotic behavior of the sum of random variables and leads to a normal distribution.

EVT has already been used to estimate very high quantiles for electrical systems (see Westerlund and Naim (2019) and Ganger et al. (2014)). Among recent applications of the EVT in the aeronautical field, the authors of Larson and Gebre-Egziabher (2017) estimate the probability of occurrence of the position, velocity or altitude errors for the navigation systems, while Sun et al. (2017) designs the load spectrum for aircraft hydraulic pumps.

The present paper illustrates the application of EVT to aeronautic electrical systems consumption to challenge the ELA assumption approach in the nominal mode only. The approach presented below can also be applied to the failure mode. Nevertheless, the few amount of data available in this case implies a specific statistical pre-treatment and is beyond the scope of the present paper.

We have a sample of 60 000 flight hours from 18 operational aircraft that we split into 8 groups

based on conditions of use of the aircraft. One main goal of our study is to use a limited amount of aircraft records to compute probabilities beyond the observed measurements. The EVT answers this challenge by estimating extreme quantiles and right endpoints. Probabilities associated with extreme quantiles are then converted into probabilities by flight hours. Confidence intervals are built to encompass the non observed aircraft.

As each aircraft has its own configuration, the ELA value may vary. Thus we choose to estimate the maximum ratios between the electrical consumption and the theoretical maximum values given by the ELA rather than estimating the maximal electrical consumption. Applying EVT on these ratios will help us to evaluate a maximal ratio irrespective of the electrical aircraft configuration.

First, we apply EVT to each group separately. Then we compare the results between the different groups by using a statistical test. The null assumption is the equality of the endpoints between groups. Using our sample we do not reject the null assumption at usual error level of 5%. From this result we can suggest a generalized maximal ratio to all operational aircraft and to the future aircraft model. Multiplying the ELA values by the maximal ratio leads to adjusted ELA value that could be used for sizing future generators or adding more loads to operational aircraft.

This paper is organized as follows. Section 1.2 presents the aircraft electrical network and details the dataset used to assess the electrical network. Section 1.3 recalls the EVT procedure and the model selection method used to estimate the extreme quantiles and endpoints. Section 1.4 illustrates this model selection procedure on a given group example. It also shows the results obtained using data from the 8 groups separately and globally after testing the endpoints equality of the ratios between groups. Finally, Section 1.5 concludes the study and proposes possible extensions.

1.2 Context and data presentation

We are interested in evaluating the extreme electrical consumption with respect to the theoretical ELA value of the generators based on operational measurements.

An aircraft flight is segmented into several phases depending on the altitude and the electrical source used. In our study we consider the flight phase, i.e. where the landing gear is no longer compressed and the altitude is greater than 1.500 feet, and the onground phase and we first analyse these phases separately.

1.2.1 Aircraft electrical network

Different electrical sources power the electrical network of an aircraft:

- AC (Alternating Current) generators are supplied by the engines. Depending on the aircraft

Table 1.2.1 – Percentage of acceptable overload for an AC generator

	under 5 sec	**under5 min**	**> 5 min**
AC loads	160% to 183%	120% to 125%	100%

Table 1.2.2 – Groups description. # stands for the quantity available

Group	# of aircraft	# of flight hours	Continent destinations
1	2	10 263	Asia
2	1	1 675	America - Europe
3	4	10 694	Europe
4	2	5 589	Asia
5	5	22 480	Asia
6	2	5 726	America - Europe - Oceania
7	1	1 825	North America
8	1	1 793	Europe - North America
Total	**18**	**60 045**	-

family, the number of AC generators is two or four. Each generator has a capacity of 90-100 Kilo-Volt-Ampere (KVA).

- APU Generator (Auxiliary Power Unit) is an additional generator that supplies energy. It is used during the onground phase and as a backup in the flight phase to replace one or more AC generators at any time.
- RAT (Ram Air Turbine) is a wind turbine and a power source in case of loss of all electrical sources.

In this paper, we focus the analysis on one of the AC generators.

The generators can support an overload that depends on the load duration. For the AC generator, the percentages of acceptable overload are shown in Table 1.2.1. The loads are classified as intermittent or permanent: the loads with a duration less than 5 minutes are called intermittent loads; otherwise, they are called permanent loads. In what follows we focus the analysis on the permanent loads only. Moreover, when there is no failure, the electrical network is in the nominal mode and we consider this mode only.

1.2.2 Data details

We have 8 groups for which we consider 18 operational low-cost aircraft from the same family. Their characteristics are given in Table 1.2.2.

For each aircraft, we observe at every second the ratio defined by the electrical consumption divided by the maximal electrical load given by the ELA for the corresponding aircraft and phase.

Let Y be a random variable which represents these ratios. The ratios are expressed in percentage but this has no impact on the EVT analysis.

We split the observations into the flight phase and onground phase and independently apply the EVT to each of the two phases.

To remove the intermittent loads, we average Y in a time window of length T by

$$X_k = \frac{1}{T}\sum_{i=1}^{T} Y_{(k-1)T+i}, \quad k \in \{1, \dots, \tau\} \tag{1.2.1}$$

where $\tau = \lfloor n/T \rfloor$ and $\lfloor \cdot \rfloor$ denotes the floor part function. The i.i.d. variables X_k distributed as a variable X are positive and can be greater than 1 if the consumption exceeds the ELA value. On top of that, a special load that generates high peaks for less than 200 milliseconds is removed.

We apply the EVT on these datasets to calculate Q_p the $(1-p)$-quantile associated to a small probability p, i.e. such that $P(X > Q_p) = p$, and the right endpoint x^* of the distribution support. The endpoint is defined by $x^* := \sup\{x : P(X \leq x) < 1\}$ and can be finite or not. It corresponds to the 1-quantile and we have $P(X > x^*) = 0$.

1.3 Extreme value theory reminder

EVT is widely used in applied fields such as hydrology, meteorology and insurance (see Castillo et al. (2005)). The objective is to estimate the probability distributions of the maxima and compute the probabilities associated with rare events.

In this paper, we want to estimate extreme quantiles and endpoint for the observed ratios $x_1, \dots, x_n$ which are considered as realizations of i.i.d. random variables $X_1, \dots, X_n$ with distribution function F. Let Q_p be the $(1-p)$-quantile and x^* the right endpoint of F. Since

$$\begin{aligned} \mathbb{P}(\max(X_1, \dots, X_n) \leq x) &= \mathbb{P}(X_1 \leq x, \dots, X_n \leq x) \\ &= F^n(x), \end{aligned}$$

$\max(X_1, \dots, X_n)$ converges in probability to x^* as n tends to infinity. To obtain a nondegenerate limit distribution we need to normalize $\max(X_1, \dots, X_n)$. To this end, we assume that there exist deterministic sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$\frac{\max(X_1, \dots, X_n) - b_n}{a_n}$$

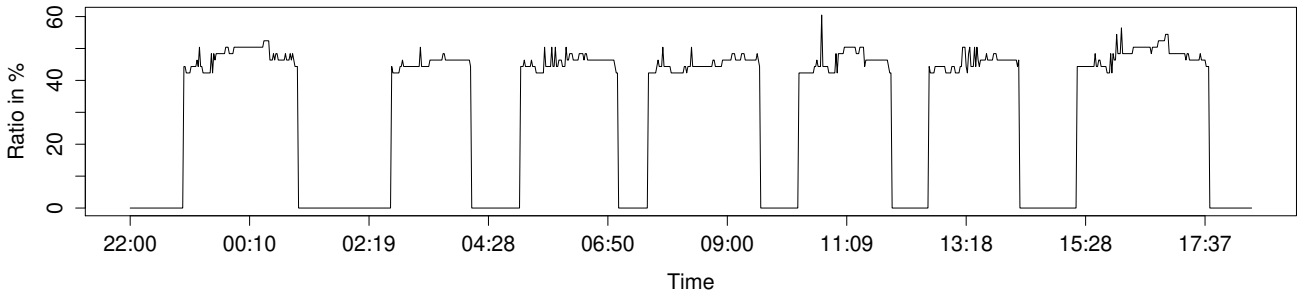

Figure 1.3-2 – Record of X for one day for a given aircraft; 7 flights were observed during this day

has a nondegenerate limit distribution as $n \to \infty$ given by

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x). \tag{1.3.1}$$

G is called extreme value cumulative distribution function and F is in the domain of attraction of G.

The previous assumption is fulfilled under regularity assumption on right endpoint of F. It can be checked for many absolutely continuous distribution functions such as uniform on an interval, normal, log-normal, gamma, beta, etc. (see details in Embrechts et al. (2013), pages 153-157).

EVT is a powerful statistical asymptotic theory that allows us to calculate extreme quantiles and endpoints without parametric assumptions on the distribution F of the data. Thanks to EVT we get a parametrized extreme distribution G. The parameters of G can be estimated using statistical methods such as the maximum likelihood or the moment method as discussed in El Adlouni et al. (2007).

The EVT is usually divided into two main approaches. The first approach is the Generalized Extreme Value (GEV) based on the study of the asymptotic distribution of a series of maxima. Under some conditions, this distribution is known to converge to Gumbel, Fréchet, or Weibull distributions. The second approach is the Generalized Pareto distribution (GPD) based on the study of the distribution of excess over a given high threshold.

The two approaches can be used to build an extreme value model for maxima and estimate the parameters. In the GEV approach the selection of the blocks size is a difficult task in practice. From our experience on the flight series data (see Figure 1.3-2), the results strongly depend on the block size and flight length, which makes the fitting difficult. This approach is more adapted to an uninterrupted series of data but is not relevant for flight data. Therefore, we only focus on the GPD approach which better captures all the maxima but recall both approaches in what follows.

1.3.1 Generalized Extreme Value approach

The GEV approach consists in dividing the series into non overlapping blocks of identical lengths and taking the maximum of each block. Let $X_1, \ldots, X_n, \ldots$ be i.i.d. random variables with unknown cumulative distribution. We define a block maximum

$$M_n = \max\{X_1, \ldots, X_n\}$$

as the observed maximum of the process over n time units. If n is the number of observations in one hour, then M_n corresponds to the maximum over one hour.

As stated in Coles et al. (2001), the asymptotic cumulative distribution function of block maximum M_n is given by

$$H(x) = \exp\left\{-(1+\xi\,\frac{x-\mu}{\sigma})^{-1/\xi}\right\}$$

where $1+\xi\,\dfrac{x-\mu}{\sigma} > 0$. The parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ correspond to location and scale, respectively. The third parameter $\xi \in \mathbb{R}$ is a shape parameter, which corresponds to the thickness of the tail of the distribution:

- $\xi > 0$ corresponds to the heavy-tailed case, and the corresponding distribution converges to Fréchet;
- $\xi = 0$ corresponds to the light-tailed case, and the corresponding distribution converges to Gumbel;
- $\xi < 0$ corresponds to the short-tailed case, and the corresponding distribution converges to Weibull.

The asymptotic distribution of the maximum is always one of these three distributions regardless of the original distribution. The asymptotic distribution of the maximum can be estimated assuming condition (1.3.1) but without any parametric assumptions on the distribution of the observations.

1.3.2 Generalized Pareto distribution approach

The GPD approach consists in selecting a given (sufficiently high) threshold and considering the observations that exceed this threshold. Let $(X_1, \ldots, X_n)$ be a sequence of independent random variables with identical distribution as X that satisfies condition (1.3.1). The random variables $X_i - u$, for $i \in \{1, \ldots, n\}$, are the exceedances over threshold u if this threshold has been exceeded.

For some μ, $\sigma > 0$ and ξ, for u sufficiently large, the cumulative distribution function of $X - u$ conditional on $X > u$ can be approximated by the distribution

$$H(x) = \begin{cases} 1 - \left(1 + \xi \dfrac{x}{\beta}\right)^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - \exp\left(-\dfrac{x}{\beta}\right) & \text{if } \xi = 0, \end{cases}$$

where $x > 0$, and $\beta = \sigma + \xi(u - \mu) > 0$ is the reparametrized scale.

Note that multiplying the random variable by a positive constant c keeps the ξ parameter unchanged while β is multiplied by c. This means that EVT is equivariant by scale transformation. Estimation of the parameters μ, σ and ξ, of the extreme quantiles and of the endpoint of the distribution F, together with their confidence intervals are derived from an asymptotic framework where u is replaced by a sequence of upper order statistics depending on n (see De Haan and Ferreira (2006) for technical details). In order to use these asymptotic results in practice, we have to ensure that the number of observations n is large but also that the ratio between n_u, the number of observations larger than u, and n is small (see Einmahl et al. (2019) for a detailed application).

The threshold selection involves balancing bias and variance. The threshold u must be sufficiently high to ensure that the asymptotic underlying the GPD approximation is reliable and thus reduce the bias. However, a reduced sample size for high thresholds increases the variance of the parameter estimators.

As discussed in Castillo et al. (2005), the common graphical diagnostics for threshold selection are the mean residual life, the threshold stability plots and the fitting diagnostic plots. These plots are described below with some guide-lines to use them for threshold selection:

- Mean residual life plot: the empirical mean of the exceedances above threshold u is plotted against u. Above threshold u_0, where the generalized Pareto distribution provides a valid approximation to the excess distribution, the mean residual life plot should be approximately linear in u.

- Threshold stability plots: ξ and β are plotted against a range of thresholds u. For u_0 selected using the mean residual life plot, we look at the stability of the parameter estimates for values of $u > u_0$ and possibly refine the choice of the threshold.

- Fitting diagnostic plots: the Probability-Probability plot and Quantile-Quantile plots, which are named PP-plot and QQ-plot, respectively, are the usual diagnostics tools. If the model fits the data, the points pattern should exhibit a 45-degree straight line for both plots. Once the threshold is selected using the mean residual life and threshold stability plots, the PP and QQ-plots are used to validate our choice.

We propose to estimate the GPD parameters using the maximum likelihood method. The log-likelihood function is given by

$$l(\xi,\beta) = \begin{cases} -n\log(\beta) - \left(\frac{1}{\xi}+1\right)\sum_{i=1}^{n}\log\left(1-\xi\,\frac{x_i}{\beta}\right), \\ \qquad \text{if } \xi \neq 0, \\ -n\log(\beta) - \frac{1}{\beta}\sum_{i=1}^{n} x_i, \\ \qquad \text{if } \xi = 0. \end{cases}$$

In practice, the values $\hat{\xi}$ and $\hat{\beta}$ that maximize $l(\xi,\beta)$ are found by using a gradient descent method (see (El Adlouni et al., 2007)).

Let X be a random variable that follows a GPD(ξ,β), the quantile Q_p is estimated by

$$\hat{Q}_p = \begin{cases} u + \frac{\hat{\beta}}{\hat{\xi}}\left[\left(\frac{n_u}{np}\right)^{\hat{\xi}} - 1\right], & \text{if } \hat{\xi} \neq 0, \\ u + \hat{\beta}\log\left(\frac{n_u}{np}\right), & \text{if } \hat{\xi} = 0. \end{cases} \tag{1.3.2}$$

It is possible to build a $(1-\alpha)$ asymptotic confidence interval (CI) for $\hat{Q}_p$ (see page 150 of De Haan and Ferreira (2006)). The upper confidence interval (UCI) limit is given by

$$Q_p < \hat{Q}_p + Z_{\alpha/2}\,\hat{\beta}\,q_{\hat{\xi}}\left(\frac{n_u}{np}\right)\sqrt{\frac{\mathrm{Var}(\hat{\xi})}{n_u}}, \tag{1.3.3}$$

where $Z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution, an approximation of q_ξ for large t (see De Haan and Ferreira (2006) page 135) is given by

$$q_\xi(t) \approx \begin{cases} t^\xi\,\log t/\xi, & \text{if} \quad \xi > 0, \\ (\log t)^2/2, & \text{if} \quad \xi = 0, \\ 1/\xi^2, & \text{if} \quad \xi < 0, \end{cases}$$

and $\mathrm{Var}(\hat{\xi})$ is the variance of $\hat{\xi}$ defined by

$$\begin{cases} (1+\xi)^2, & \text{if} \quad \xi \geq 0, \\ 1+4\xi+5\xi^2+2\xi^3+2\xi^4, & \text{if} \quad \xi < 0. \end{cases}$$

Let x^* be the right endpoint or the upper limit of the distribution. If the endpoint is known to be finite then $\xi < 0$ and an estimator of x^* can be calculated by letting $p \to 0$ in (1.3.2), which

leads to

$$\hat{X}^* = u - \frac{\hat{\beta}}{\hat{\xi}}, \text{ for } \hat{\xi} < 0. \tag{1.3.4}$$

Replacing q_ξ by $1/\xi^2$ in (1.3.3), we get $(1-\alpha)$ one sided asymptotic CI

$$x^* < \hat{X}^* + Z_\alpha \frac{\hat{\beta}}{\hat{\xi}^2} \sqrt{\frac{\text{Var}(\hat{\xi})}{n_u}}, \tag{1.3.5}$$

where Z_α is the $(1-\alpha)$ quantile of the standard normal distribution. In the next section α is called the error level.

The upper confidence interval values for the quantiles of order p and the endpoint are based on approximations that are valid under certain conditions. These conditions involve that the total number of observations n together with the number of observations that exceed the threshold u are large while the proportion n_u/n is small. Moreover, concerning the UCI of an extreme quantile, the probability p has to be small enough so that np/n_u is small but not too small in order to have a small value for $|log(np)|/\sqrt{n_u}$ (see Remark 4.3.4, page 135 in De Haan and Ferreira (2006)). Interested readers could find more details about the CI building in Chapter 4 of De Haan and Ferreira (2006).

1.4 Extreme value application on electrical loads

1.4.1 Illustration of the GPD procedure for one group

In this section, we select one group, apply the GPD approach on the data and compute upper confidence interval values for extreme quantiles and endpoint.

The group under study was observed during more than 10 000 flights between 2016 and 2018. To illustrate the results of the methodology, we select one generator in the permanent mode during the onground phase. For each flight, we apply a mean time window of $T = 150$ seconds as detailed in Equation (1.2.1). We apply the GPD approach using the package *extRemes* Gilleland et al. (2016) in the R software with the maximum likelihood estimation method.

In the first step, we set threshold u using the graphical diagnostics from section 1.3.2. The mean residual life plot is represented by a solid line in Figure 1.4-3. We look for a linear trend at the extreme right of this curve. For u between 50% and 63%, the data exhibit such a linear trend. This choice is refined using Figure 1.4-4, where we focus on u between 50% and 63%. According to these plots, $\hat{\beta}$ and $\hat{\xi}$ reach stability when $u > 57.5\%$, which indicates that the assumption of GPD is reasonable for $u \in [57.5\%, 60\%]$.

Table 1.4.1 gives the maximum likelihood estimates of $\hat{\beta}$ and $\hat{\xi}$ and confirms the stability of the

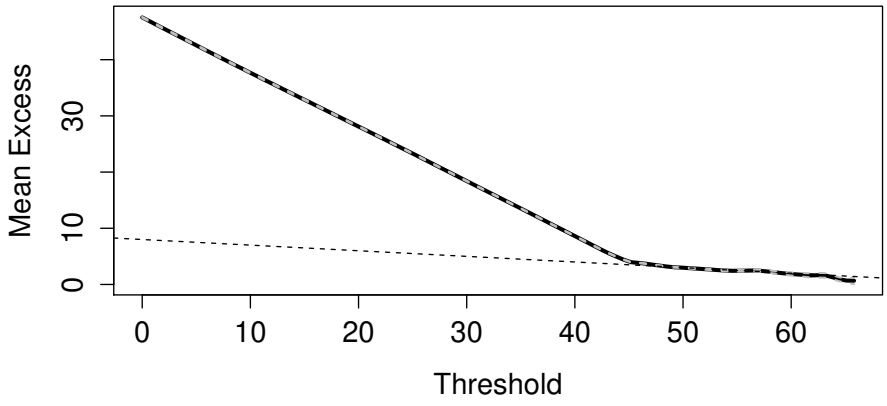

Figure 1.4-3 – Mean residual life plot. We plot u against the mean excess for a range of threshold values. A linear trend is observed for $u > 50\%$ represented by the dashed line

Table 1.4.1 – Maximum likelihood estimates of β and ξ for different thresholds u

$u(\%)$	$\hat{\beta}$	$\hat{\xi}$
57.5	3	-0.3
58	2.9	-0.2
58.5	2.7	-0.2
59	2.3	-0.2
59.5	2.3	-0.2
60	2.2	-0.2

estimates for this range of values. Then, we set $u = 59.5\%$ and check whether the model fits the data by using the fitting diagnostic PP-plot and QQ-plot in Figures 1.4-5 and 1.4-6, respectively.

In both Figures 1.4-5 and 1.4-6, the point pattern exhibits a 45-degree linear trend. So the GPD assumption appears reasonable for $u = 59.5\%$ and we obtain $n_u = 150$ from $n = 18\,319$.

To align with the safety assumption study, we have to convert our probabilities into probabilities by flight hour. In our case, we recall that the data are preprocessed by taking the mean of the consumption during a time window of $T = 150$ seconds (see Section 1.2). Therefore, we have 24 observations per hour.

Let $X_1, ..., X_{24}$ be the variables observed during a given hour, we want to compute the probability that their maximum is above the quantile Q_p. For a given probability p to exceed Q_p during a period of length T and assuming that $X_1, ..., X_{24}$ are i.i.d. with the same distribution as X, we can write

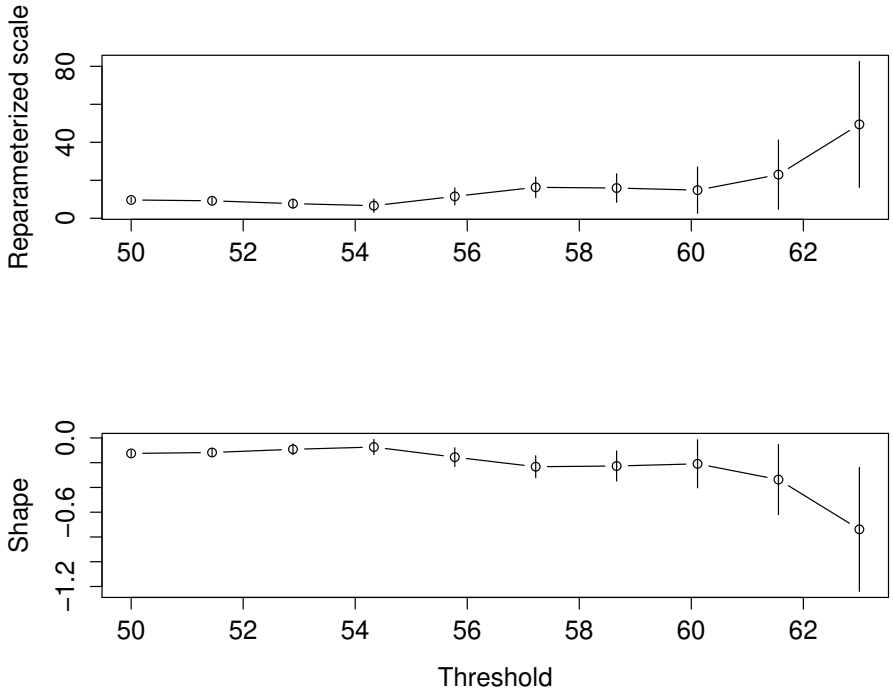

Figure 1.4-4 – Threshold stability plots for a threshold between 50% and 63% (top plot for β and bottom plot for ξ). For each value of u the vertical bar represents the confidence interval of the estimators. Stability of estimators is observed for $u \in [57.5\%, 60\%]$

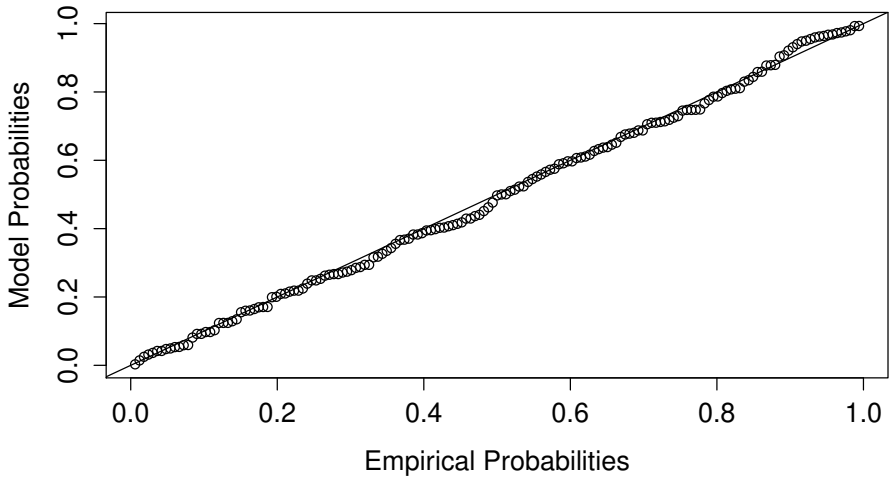

Figure 1.4-5 – PP-plot obtained from fitting the GPD using the maximum likelihood method for $u = 59.5\%$. The point pattern falls along the 45-degree line represented by the black line

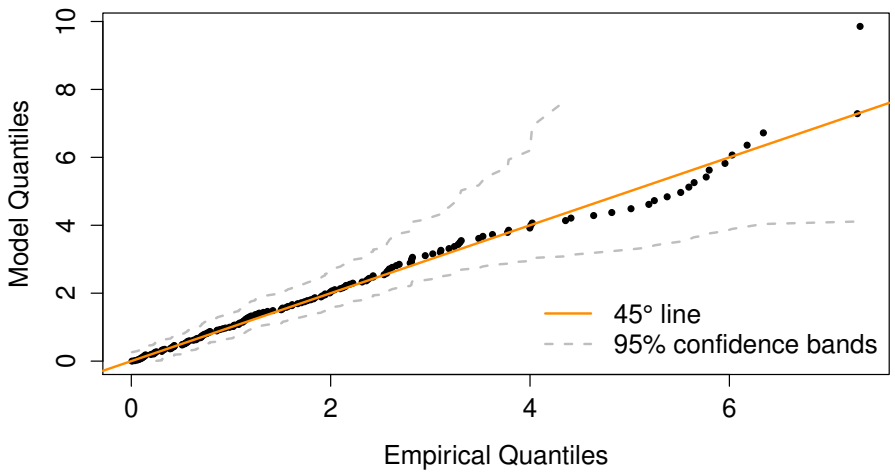

Figure 1.4-6 – QQ-plot obtained from fitting the GPD using the maximum likelihood method for $u = 59.5\%$. The point pattern falls along the 45-degree line. The dashed lines represent the 95% confidence bands based on the Kolmogorov-Smirnov statistics

Table 1.4.2 – Quantile estimation for different values of P_{hour}

P_{hour}	p	Q_p
10^{-3}	10^{-5}	67.1
10^{-5}	10^{-7}	69.4
10^{-7}	10^{-9}	70.3
10^{-9}	10^{-11}	70.7
10^{-12}	10^{-14}	70.9

$$\begin{aligned}\mathbb{P}\left(\max_i X_i > Q_p\right) &= 1 - \mathbb{P}\left(\max_i X_i \leq Q_p\right) \\ &= 1 - \mathbb{P}\left(X \leq Q_p\right)^{24} \\ &= 1 - \left[1 - \mathbb{P}\left(X > Q_p\right)\right]^{24} \end{aligned} \tag{1.4.1}$$

Let P_{hour} be the probability to exceed Q_p in one hour. Then Equation (1.4.1) becomes $P_{hour} = 1 - (1-p)^{24}$, and we can compute p for a target probability P_{hour}. Table 1.4.2 shows the results obtained using Equations (1.3.2) and (1.4.1) to estimate quantiles associated to the target probabilities.

From Table 1.4.2 we select the result corresponding to $P_{hour} = 10^{-7}$ to respect the aeronautic safety procedure and not increase the probability of losing one generator.

At the probability 10^{-7} by flight hour, the maximum ratio for the selected generator is 70.3%. Using the results from Equation (1.3.3) we build UCI at error levels $\alpha = 5 \times 10^{-2}, 10^{-2}, 10^{-3}, 10^{-5}$,

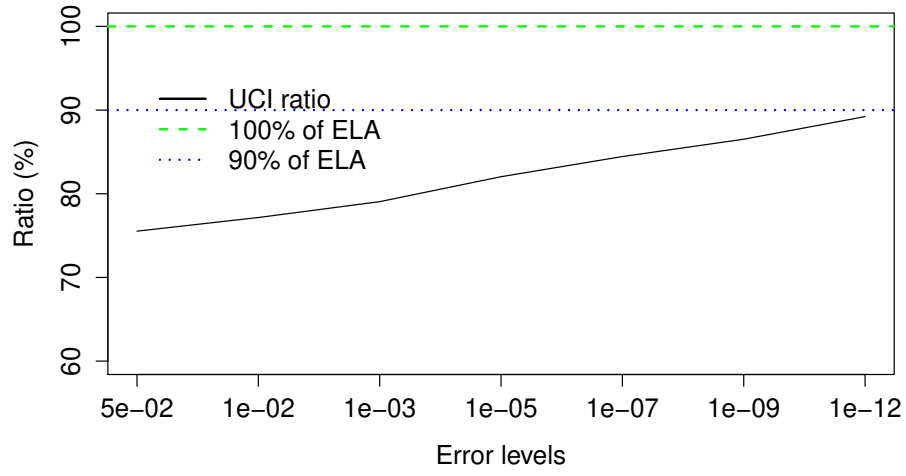

Figure 1.4-7 – UCI for the quantile associated to the probability 10^{-7} by flight hour with respect to the error levels. The green dashed line (resp. blue dotted line) represent 100% (resp. 90%) of the ELA value

$10^{-7}, 10^{-9}$, 10^{-12} and plot these UCI with respect to the error levels. Figure 1.4-7 shows a trend from 70.2% to 88.3%.

From Table 1.4.1 we see that $\hat{\xi}$ is always negative and so we can assume that the endpoint exists and, from Equation (1.3.4), is estimated at 71%. Using Equation (1.3.5) we can build a CI around the endpoint estimate. Figure 1.4-8 gives the endpoint CI with respect to the error levels $\alpha = 5 \times 10^{-2}, 10^{-2}, 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}, 10^{-12}$. It shows a trend between 75% and 90%.

The results from the quantile at $P_{hour} = 10^{-7}$ and the endpoint are close. For the group under study, with a reasonable risk error ($\alpha = 10^{-3}$) and to remain in accordance with the aeronautic safety procedures ($P_{hour} = 10^{-7}$), we can consider a ratio of 80% which means that the ELA is overestimating the electrical network by 20% with an error level of 10^{-3} for this group.

Concerning the assumptions advocated at the end of Section 1.3.2, most are clearly fulfilled in our context, namely that $n = 18\,319$ and $n_u = 150$ are large while $n_u/n = 8 \times 10^{-3}$ and $np/n_u = 5 \times 10^{-5}$ are small. It is not as clear when it comes to the assumption that $|log(np)|/\sqrt{n_u}$ is small since it equals 0.77. It means that the extrapolation should not be pushed further and results concerning the UCI of extreme quantiles with smaller P_{hour} than 10^{-7} may not be valid anymore.

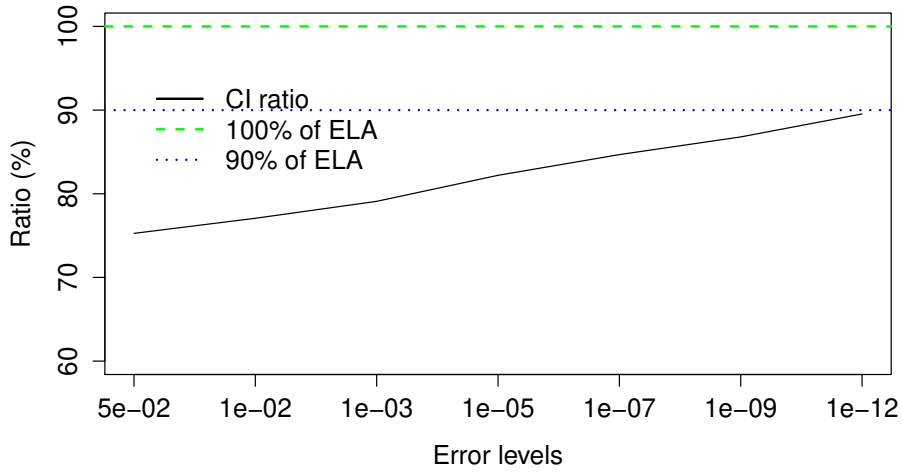

Figure 1.4-8 – CI for the endpoint with respect to the error levels. The green dashed line (resp. blue dotted line) represents 100% (resp. 90%) of the ELA value

Table 1.4.3 – Maximum likelihood estimates $\hat{\beta}$ and $\hat{\xi}$ by group for the flight and onground phases

Group	Flight phase					Onground phase				
	n	n_u	n_u/n	$\hat{\beta}$	$\hat{\xi}$	n	n_u	n_u/n	$\hat{\beta}$	$\hat{\xi}$
1	227 988	316	0.001	2.23	-0.18	18 319	150	0.008	2.47	-0.24
2	35 504	150	0.004	3.46	-0.41	4 701	150	0.032	1.96	-0.24
3	232 296	150	0.001	2.2	-0.33	24 349	13	0.001	2.75	-0.34
4	113 787	200	0.002	1.64	-0.16	20 355	637	0.031	2.19	-0.16
5	455 263	430	0.001	1.91	-0.23	84 267	26	0.000	3.64	-0.44
6	123 430	600	0.005	3.19	-0.24	13 987	500	0.036	2.05	-0.21
7	38 063	150	0.004	2.1	-0.3	5 728	80	0.014	1.76	-0.35
8	40 104	120	0.003	3.15	-0.28	2 935	50	0.017	1.19	-0.22

1.4.2 Global results

Using the EVT on the sampled groups we want to demonstrate that the ELA is overestimating maximal consumption for all groups. For that, we apply separately the same procedure to the 8 groups for the flight and onground phases to estimate extreme quantile, endpoint and their confidence intervals.

We use the same procedure as described in Section 1.4.1 to set the threshold and fit the GDP. Table 1.4.3 shows the parameter estimates for each group by phase. We see that the number of observations for the onground phase is smaller than for the flight phase which is coherent given the length of the two phases. All $\hat{\xi}$ are negative which implies a finite endpoint for all groups in both phases.

Table 1.4.4 – Quantiles associated to the probability 10^{-7} by flight hour and its UCI at error level of 10^{-3} by group for the flight and onground phases

Group	Flight phase		Onground phase	
	$\hat{X}_{10^{-7}}$	$\text{UCI}_{10^{-3}}$	$\hat{X}_{10^{-7}}$	$\text{UCI}_{10^{-3}}$
1	67.3	75.2	69.5	75.7
2	70.3	72.2	68.5	73.5
3	69.5	71.8	74.2	83.0
4	65.3	75.4	69.5	77.2
5	69.1	72.2	72.6	76.8
6	70.9	75.2	72.1	76.0
7	67.4	70.4	70.3	72.3
8	71.9	77.7	66.0	73.0

Table 1.4.5 – Endpoints and its CI at error level 10^{-3} by group for the flight and onground phases

Group	Flight phase		Onground phase	
	$\hat{X}^*$	$\text{CI}_{10^{-3}}$	$\hat{X}^*$	$\text{CI}_{10^{-3}}$
1	68.5	75.9	69.8	75.6
2	70.3	72.1	68.7	73.4
3	69.7	71.8	74.4	82.6
4	66.6	76.1	70.6	77.8
5	69.6	72.5	72.7	76.6
6	71.4	75.5	72.4	76.1
7	67.5	70.3	70.3	72.2
8	72.2	77.5	66.2	72.8

To compare the maximal electrical consumption between groups we need to compute the extreme quantiles and endpoint ratios by groups. Let $\hat{Q}_{10^{-7}}$ be the estimated extreme quantile associated to $P_{hour} = 10^{-7}$ and $\text{UCI}_{10^{-3}}$ its UCI at error level $\alpha = 10^{-3}$. Let $\text{CI}_{10^{-3}}$ be the CI at error level $\alpha = 10^{-3}$ for the estimated endpoint $\hat{X}^*$. The quantiles and endpoints estimates are given in Tables 1.4.4 and 1.4.5. Concerning the assumptions advocated at the end of Section 1.3.2, we can see that not all of them are fulfilled for all groups. The size n is large and the ratios n_u/n and np/n_u are small in all situations. But the size n_u is quite small and $|log(np)|/\sqrt{n_u}$ is quite large for the groups 3, 5, 7 and 8 for the onground phase. It means that the results concerning the UCI of the quantiles and the endpoints for these three groups during the onground phase have to be interpreted with caution. It also justifies the interest of gathering the different groups and phases if the results are sufficiently similar.

We observe that $\hat{Q}_{10^{-7}}$ and $\hat{X}^*$ are close. This can be explained by the fact that we are computing quantiles associated to $p = 10^{-9}$ to get the target probability 10^{-7} by flight hours and this probability is so small that we almost reach the endpoint. We see that the CI of the endpoint

ratios by groups are aligned in a range of $70\% - 83\%$ which confirms our assumptions that the ELA overestimates the electrical consumption for permanent loads in nominal mode for the observed groups.

The largest endpoint ratio observed is 78% and 83% respectively for flight and onground phases but the ratio varies from one group to another. The final aim of this work is to generalize the observed ratio to all operational aircraft and to size the future aircraft generator. For that, we need to test if the endpoints can be considered the same for the different groups.

To this end we use an asymptotic chi-square test developed in Einmahl et al. (2019). This test checks the equality of the endpoints for independent random samples. We apply this test to check the equality of the group endpoints. We can consider that the assumption of independence between groups is satisfied as the electrical consumption of one group does not depend on the consumption of another. Let x_j^* be the endpoint of the j^{th} group with $j = 1, \ldots, 8$. We consider the following hypotheses

$$\begin{cases} H_0 : x_1^* = \cdots = x_8^* \\ H_1 : \text{the } x_j^* \text{ are not all equal.} \end{cases}$$

The test statistic is

$$S = d \sum_{j=1}^{8} r_j (\hat{X}_j^* - \tilde{X})^2$$

where $\tilde{X} = \sum_{j=1}^{8} r_j \hat{X}_j^*$, with $r_j = \dfrac{d_j}{d}$, $d = \sum_{j=1}^{8} d_j$, $d_j = \dfrac{n_u^j}{\hat{\beta}_j^2 \, \tau(\hat{\xi}_j^2)}$ and $\tau(\hat{\xi}_j)^2 = 2 + 2\xi_j^{-1} + 5\xi_j^{-2} + 4\xi_j^{-3} + \xi_j^{-4}$, where n_u^j (resp. $\hat{\xi}_j$ and $\hat{\beta}_j$) are the number of observations that exceed threshold u (resp. the shape and the scale estimators) for group j.

Under H_0, Einmahl et al. (2019) demonstrates that the test statistic S follows a chi-square distribution with 7 degrees of freedom. We reject H_0 at level α if $S > q_{\chi_7^2(1-\alpha)}$ where $q_{\chi_7^2(1-\alpha)}$ stand for the $(1-\alpha)$-quantile of the chi-square distribution with 7 degrees of freedom.

The result of this statistical test is given in Table 1.4.6. The p-values for both phases are greater than 0.05 hence the hypothesis that the endpoints are equal is not rejected with a 5% risk error.

We do not reject that group endpoints are equal (for both phases) which means that the largest possible value of the maximum of electrical network consumption divided by the ELA value does not depend on the groups. This result can also be deduced from Figures 1.4-9 and 1.4-10 where the endpoints estimates are represented by dots and the corresponding CI by dashed bars. These figures are graphical representations in connection with the chi-square test results and help us to check the equality of endpoints. We confirm graphically the equality of endpoints for both phases

Table 1.4.6 – Chi-square test for groups endpoint equality in flight and onground phases

Group	$\hat{X}^*$ flight	$\hat{X}^*$ onground
1	68.5	69.8
2	70.3	68.7
3	69.7	74.4
4	66.6	70.6
5	69.6	72.7
6	71.4	72.4
7	67.5	70.3
8	72.2	66.2
S	11.7	13.1
p-value	0.11	0.07

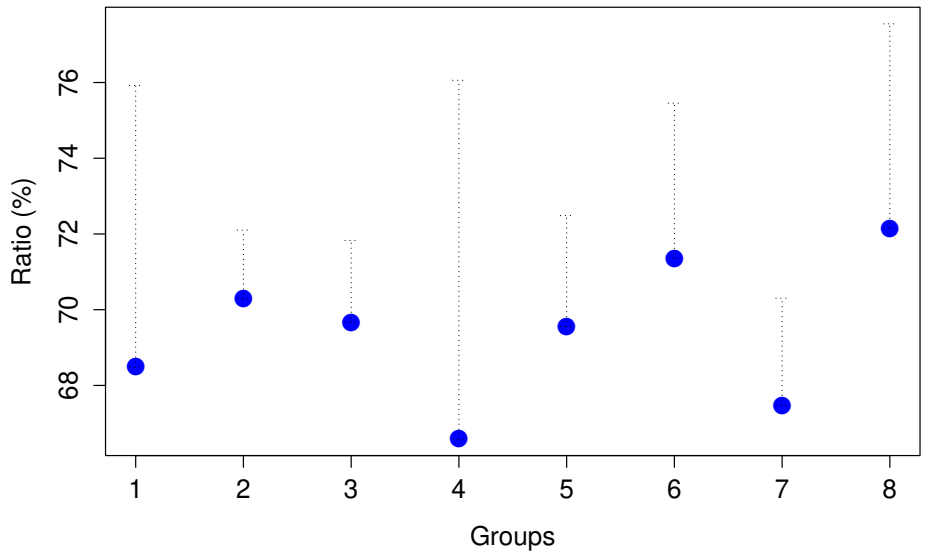

Figure 1.4-9 – Endpoints and their CI by group for the flight phase represented by the dashed bars

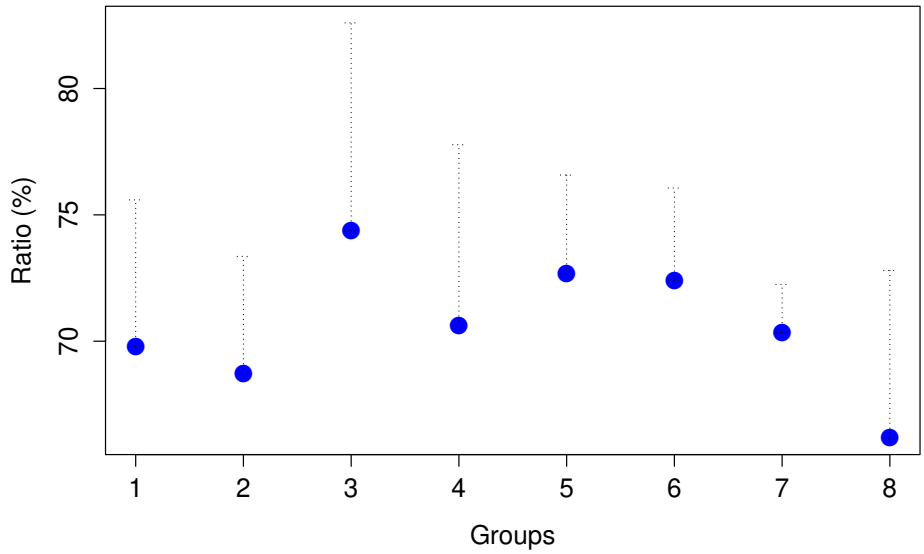

Figure 1.4-10 – Endpoints and their CI by group for the onground phase represented by the dashed bars

Table 1.4.7 – Parameters, quantiles, endpoint estimates and their confidence interval for gathered group by phase

Phase	n	n_u	n_u/n	$\hat{\beta}$	$\hat{\xi}$	$\hat{X}_{10^{-7}}$	$\text{UCI}_{10^{-3}}$	$\hat{X}^*$	$\text{CI}_{10^{-3}}$
Flight	1 266 435	335	0.000	1.84	-0.24	71.5	74.7	71.6	74.6
Onground	174 640	120	0.001	1.45	-0.2	73.3	79.7	73.5	79.6

since the CI intersect with each other on the two figures.

As the endpoints equality test suggests that there is no effect of the group on the estimated ratio, we gather all groups and estimate a global ratio taking into account all groups. We apply the EVT separately to the flight and the onground phases. The parameters, extreme quantiles, endpoints estimates and their CI are given in Table 1.4.7. It shows that we still have a negative $\hat{\xi}$ and thus a finite endpoint. The ratio estimates of extreme quantile and endpoint are close and around 75% for the flight phase and around 80% for the onground phase. Comparing to the ratios found in Table 1.4.5 the results are aligned.

To go further in generalizing this ratio and since the endpoints for flight and onground phase are close we check if the endpoints are equal. Table 1.4.8 provides the results of the chi-square test of endpoint equality between the flight and the onground phases. The test illustrates that we cannot reject the equality of endpoints at 5% error level and thus the estimated ratio can be considered as independent of the phase.

From this result we gather also the two phases and apply the EVT on the gathered groups with no distinction between flight and onground phases. Table 1.4.9 shows the maximum likelihood estimates of the parameters β and ξ for the gathered groups and phases, we still have $\hat{\xi} < 0$ and

Table 1.4.8 – Chi-square test for phases endpoint equality

Phase	$\hat{X}^*$
Flight	71.6
Onground	73.5
S	0.8
p-value	0.381

Table 1.4.9 – Parameters estimates associated to the gathered groups and phases

n	n_u	n_u/n	$\hat{\beta}$	$\hat{\xi}$
1 441 076	500	0.000	1.5	-0.13

thus consider a finite endpoint.

The extreme quantile is estimated at 72.9% and the endpoint at 74.3%. The UCI and CI are given in Figures 1.4-11 and 1.4-12 where we vary the error levels $\alpha = 5 \times 10^{-2}, 10^{-2}, 10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}, 10^{-12}$ and plot the UCI of extreme quantile and CI of the endpoint with respect to the error level. As could be expected, we observe an increasing trend for extreme quantile and endpoint ratios. They both vary from 75% to 87%. We see that for the error level $\alpha = 10^{-3}$ we have a ratio of 80% which is in line with the previous results.

In all applications of the EVT, by groups and on gathered data, we get a maximal ratio of 80% for an error level 10^{-3}. From these results we can consider a ratio of 80% for the generator with permanent loads in nominal mode.

1.5 Conclusion

In this paper, we use the extreme value theory to estimate extreme ratios associated to probability 10^{-7} by flight hour and endpoint ratios, we also build confidence intervals at error level 10^{-3} to check whether the ELA overestimates the maximal consumption. We detail the statistical procedure for permanent loads of a generator in the nominal mode for a specific group. Then, we apply the EVT to 8 groups and demonstrate that the largest ratio is around 83% for the permanent loads in the nominal mode.

To generalize this gap to all operational aircraft and to size the future aircraft generators, we do an asymptotic chi-square test to check that the group endpoints are equal. The endpoints equality is not rejected for both phases which means that there is no group effect on the ratio endpoint. Then we gather all groups to estimate extreme quantiles and endpoint ratios for each of the two phases and we end up with a ratio of 75% for flight phase and 80% for the onground phase. To obtain a global ratio, we check if there is a difference between the flight and onground phases using

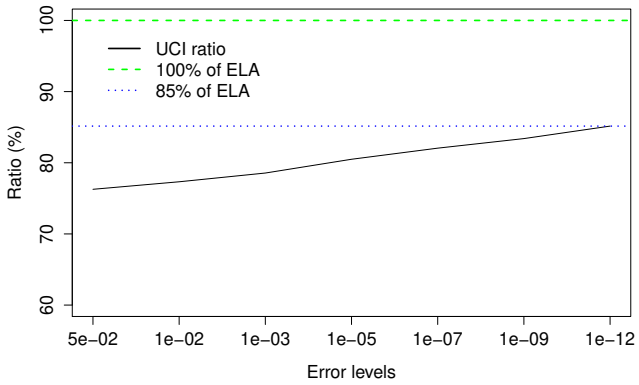

Figure 1.4-11 – UCI for the quantile associated to the probability 10^{-7} by flight hour with respect to the error levels for the gathered groups and phases. The dotted line represents the maximum ratio obtained

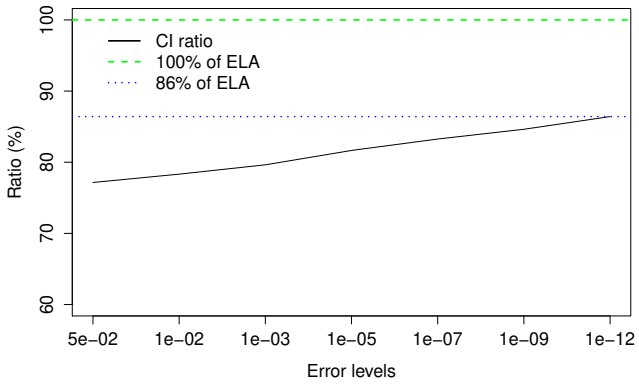

Figure 1.4-12 – CI for the endpoint with respect to the error levels for the gathered groups and phases. The dotted line represents the maximum endpoint ratio obtained

the endpoint equality test. Again the equality assumption is not rejected and after gathering the two phases, we obtain an endpoint ratio of 80%.

Using a statistical approach, we quantify how much the ELA overestimates the maximal electrical consumption of the generator. For instance, with an error level of 10^{-3} for permanent loads in the nominal mode, our study leads to an excess of 20% for the considered generator.

However, the study only relies on permanent loads in the nominal mode for low-cost aircraft. To complete the electrical network assessment, we need to incorporate also non low-cost aircraft in our analysis and extend the study to the intermittent loads and failure modes. In particular, future work should focus on the degraded mode (loss of generators) to size the generators.

Conclusion

In this part, we applied the EVT to 8 different groups of aircraft to estimate the maximal ratio between the electrical consumption and the maximal ELA value. We applied the EVT separately by group and by generator in the flight and onground phases. We estimated extreme quantiles, endpoints and their upper confidence interval value. We applied EVT for the permanent loads in the nominal mode and we ended up with an endpoint ratio of 80% with error level 10^{-3}. Using a statistical approach, we confirm that the ELA is overestimating the generator capacity by 20% for the permanent loads in the nominal mode. This result should be pushed further by adding more aircraft as we are covering less than 2% of the airbus fleet.

Part II

Oil temperature prediction of aeronautical electrical generator

Sommaire

Introduction

The oil circuit cools the electrical generator and its temperature may be used as a measurement of proper functioning of the generator. An overheating may generate a faulty operation and ultimately a generator failure. For these reasons, the prediction of the generator oil temperature helps engineers to understand and simulate the behavior of the electrical generator under extreme conditions.

To capture the history of the features, we choose a functional data representation to predict the oil temperature at a given time using the observation history of the auxiliary functions. Functional data representation is used in different fields and for several purposes like the work of Laukaitis (2008) in Finance to predict the cash flow and intensity of transactions in a credit card payment systems or the work of Ratcliffe et al. (2002) in medicine to predict the age-specific breast cancer mortality.

In this part, we use functional data analysis tools to reduce the model complexity by expanding our functions onto orthonormal bases. This step reduces the model complexity but decreases the model accuracy. On top of that, we are handling a data set with a large parameter dimension with respect to the number of observations. This leads to overfitting phenomena, characterized by a good accuracy on the initial data, called training data set, but a poor quality of the predictions based on new data, called testing data set.

To avoid the pitfall of overfitting, we need to use a regularization method. Several techniques exist, the most common ones are the Ridge and Lasso regularizations but a trending technique known as dropout regularization is arising from the neural network community. The dropout is a popular regularization method in deep learning community since the work of Srivastava et al. (2014). It allows to reduce overfitting by randomly dropping entries at training time. This can be seen as a way to regularize by adding noise to the model. We are interested in the works of Mianjy et al. (2018), Wager et al. (2013) and Arora et al. (2020) to develop an explicit expression of the dropout term added to the non regularized loss function.

In the first chapter of this part, we detail the way to use multivariate functional data to predict a scalar output. We describe the dropout technique for linear model and neural networks and compare the results of several ways to dropout on flights data set.

In the second chapter, we give an application of multivariate functional data prediction in a predictive maintenance context. We compare the most used procedures and select the best that predict the generator oil temperature and integrate this model into a process that detect abnormal behavior of the generator. This work was published in the international Global Congress on Electrical Engineering (GC-ElecEng) in September 2020.

Chapter 2

Regression prediction models in functional data framework

We are interested in the prediction of the generator oil temperature using the history of the observed features. The features are sensor data recorded each second during the flights. Functional Data Analysis (FDA) is commonly used for time series as it captures the history of the features and give more accuracy to the models.

The FDA considers each sample as function defined in the Hilbert space of square integrable functions. The FDA tools are useful for prediction since it allows us to reduce the dimension by summarizing the information on some coefficients along an orthonormal basis. Using such an expansion, we will end up with coefficients instead of functions and apply standard machine learning procedures on the resulting coefficients.

In this chapter we detail the expansion procedure of multivariate functional data. We recall the prediction procedures of neural network, random-forest and linear regression that we use on the output of the expansion procedure.

We study the effect of the dropout technique on the loss function. We write an explicit expression of the dropout regularization term for linear regression and single layer perceptron. We develop algorithms that compute stochastic gradient descents for linear regression and single layer perceptron. For each prediction procedure, we discuss its application, regularization and give some experimental results.

2.1 Functional data

The observations can be represented as curves if the explanatory variables are observed in continuous set of values (interval), e.g. time or multidimensional like images.

Let assume that our data are vectors of square integrable functions on $[0,1]$ with respect to Lebesgue measure dt. We denote the Hilbert space of such functions by $L^2 = L^2([0,1], dt)$. Hereafter, we consider the usual inner product,

$$\forall f, g \in L^2, \langle f, g \rangle_{L^2} = \int_0^1 f(t)\, g(t)\, dt.$$

Let $\{\xi_k\}_{k\in\mathbb{N}}$ be an orthonormal basis in L^2 with respect to Lebesgue measure dt. The expansion of a function $f \in L^2$ according to the orthonormal basis is given by

$$\forall t \in [0,1], f(t) = \sum_{k\in\mathbb{N}} c_k\, \xi_k(t),$$

where

$$c_k = \langle f, \xi_k \rangle_{L^2} = \int_0^1 f(t)\, \xi_k(t)\, dt.$$

For our purposes, the functions are only observed on a discrete fixed design. Thus, coefficients c_k are not available and we handle estimators $\hat{c}_k$ on a regular grid of step $\frac{1}{T}$ in $[0,1]$ given by

$$\hat{c}_k = \frac{1}{T} \sum_{i=1}^{T} f\left(\frac{i}{T}\right)\, \xi_k\left(\frac{i}{T}\right).$$

To reduce the dimension, we plan to handle a truncation by selecting the first $K \in \mathbb{N}$ coefficients of each explanatory variable which means that we have the same number of coefficients K per auxiliary function. On the selected coefficients, we apply usual machine learning procedures for regression prediction.

There are many possible orthonormal bases (Ramsay and Silverman, 2005) and we consider hereafter two common cases given by Fourier and Haar bases.

2.1.1 Expansion on Fourier basis

The Fourier expansion breaks down a L^2 function into the sum of trigonometric functions. The Fourier basis of L^2 is given by

$$\begin{aligned}
&\xi_0(t) = 1 \\
&\xi_{2k-1}(t) = \sqrt{2}\sin(2\pi k t) \\
&\xi_{2k}(t) = \sqrt{2}\cos(2\pi k t)
\end{aligned}$$

with $t \in [0,1]$ and $k \in \mathbb{N}^*$. The expansion of $f \in L^2$ on Fourier gives the coefficients

$$c_0 = \langle f, \xi_0 \rangle_{L^2} = \int_0^1 f(t)\, dt$$

$$c_{2k-1} = \langle f, \xi_{2k-1} \rangle_{L^2} = \int_0^1 \sqrt{2}\, f(t)\, \sin(2\pi k t)\, dt$$

$$c_{2k} = \langle f, \xi_{2k} \rangle_{L^2} = \int_0^1 \sqrt{2}\, f(t)\, \cos(2\pi k t)\, dt.$$

Thus, estimated coefficients are given by

$$\hat{c}_0 = \frac{1}{T} \sum_{i=1}^{T} f\left(\frac{i}{T}\right)$$

$$\hat{c}_{2k-1} = \frac{\sqrt{2}}{T} \sum_{i=1}^{T} f\left(\frac{i}{T}\right) \sin\left(2\pi k \frac{i}{T}\right)$$

$$\hat{c}_{2k} = \frac{\sqrt{2}}{T} \sum_{i=1}^{T} f\left(\frac{i}{T}\right) \cos\left(2\pi k \frac{i}{T}\right).$$

2.1.2 Expansion on Haar wavelet basis

There are two functions that play a central role in wavelet analysis, the father wavelet ψ and the mother wavelet ξ. For the sake of simplicity, we consider the Haar basis but our study remains valid for any other wavelet basis. The father function is defined by

$$\forall t \in \mathbb{R}, \quad \psi(t) = \begin{cases} 1 & \text{if} \quad 0 \leq t < 1 \\ 0 & \text{otherwise,} \end{cases}$$

and the mother function is defined by $\xi(t) = \psi(2t) - \psi(2t-1)$, which gives

$$\xi(t) = \begin{cases} 1, & \text{if} \quad 0 \leq t < 1/2 \\ -1, & \text{if} \quad 1/2 \leq t < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The Haar basis in L^2 is then given by

$$\xi_{k,j}(t) = 2^{k/2}\, \xi(2^k t - j)$$

where $k \in \mathbb{N}$ and $j \in \left\{0, 1, \ldots, 2^k - 1\right\}$. Which give us Haar coefficients

$$c_0 = \langle f, \psi \rangle_{L^2} = \int_0^1 f(t)\, \psi(t)\, dt$$

$$c_{k,j} = \langle f, \xi_{k,j} \rangle_{L^2} = \int_0^1 f(t)\, \xi_{k,j}(t)\, dt.$$

Assuming that T is a power of 2, the discretization of the Haar coefficients remains an orthonormal collection of $\mathbb{R}^T$ and we can approximate the coefficients by

$$\hat{c}_0 = \frac{1}{T} \sum_{i=1}^{T} f\left(\frac{i}{T}\right)\, \psi\left(\frac{i}{T}\right)$$

$$\hat{c}_{k,j} = \frac{1}{T} \sum_{i=1}^{T} f\left(\frac{i}{T}\right)\, \xi_{k,j}\left(\frac{i}{T}\right).$$

Otherwise, some orthonormalization procedure is needed but this is beyond the scope of this introduction.

2.2 Prediction procedures

In the sequel, we use similar notations as Arora et al. (2020). We denote matrices, vectors, scalar variables and sets by Roman capital letters, Roman small letters, small letters, and script letters, respectively (e.g. X, x, x, $\mathcal{X}$). For any vector $\mathrm{x} \in \mathbb{R}^d$, $\|\mathrm{x}\|$ represents the l_2-norm of x and $\mathrm{diag}(\mathrm{x}) \in \mathbb{R}^{d \times d}$ is the diagonal matrix whose diagonal entries are given by the entries of x. Conversely, for any matrix $\mathrm{X} \in \mathbb{R}^{d \times d}$, $\mathrm{diag}(\mathrm{X}) \in \mathbb{R}^d$ stands for the vector whose entries are the diagonal entries of X and the diagonal matrix $\mathrm{ddiag}(\mathrm{X}) \in \mathbb{R}^{d \times d}$ is defined as $\mathrm{ddiag}(\mathrm{X}) = \mathrm{diag}(\mathrm{diag}(\mathrm{X}))$. For any real function $\varphi : \mathbb{R} \to \mathbb{R}$ and any vector $\mathrm{x} \in \mathbb{R}^d$, $\varphi(\mathrm{x}) \in \mathbb{R}^d$ denotes the vector obtained by applying φ to each entry of x. For any vectors $\mathrm{x}, \mathrm{x}' \in \mathbb{R}^d$, we denote the Hadamard product by $\mathrm{x} \odot \mathrm{x}' \in \mathbb{R}^d$.

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$ denote the input and output spaces, respectively. We consider data $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ drawn i.i.d. from an unknown joint probability distribution D on $\mathcal{X} \times \mathcal{Y}$. We aim to construct a regression predictor $f_\omega : \mathcal{X} \to \mathcal{Y}$ to predict $\mathrm{y} = (y_1, \ldots, y_n)'$ solely based on the input data $\mathrm{X} = (\mathrm{x}_1, \ldots, \mathrm{x}_d)$, where $\omega \in \Omega \subset \mathbb{R}^q$ stands for some parameters to be defined. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, the quality of the predictor f_ω is quantified by the expected risk

$$L(f_\omega) = \mathbb{E}_D \left[\ell(y, f_\omega(\mathrm{x}))\right].$$

For our purpose, we focus on the quadratic loss function $\ell(y, y') = (y - y')^2$ and the associated

empirical risk

$$\widehat{L}(f_\omega) = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - f_\omega(\mathrm{x}_k)\right)^2. \tag{2.2.1}$$

The goal of this work is to study algorithms to solve the minimization problem

$$\min_{\omega\in\Omega} \widehat{L}(f_\omega). \tag{2.2.2}$$

To this end, we select common regression predictors: linear regression, neural network and random-forest and detail hereafter each algorithm.

2.2.1 Linear regression

Linear regression assumes a linear relationship between the input variables and the response variable defined by $f_\omega(\mathrm{x}) = \mathrm{w}^\top\mathrm{x} + w_0$ with $\omega = (\mathrm{w}, w_0) \in \mathbb{R}^d \times \mathbb{R}$. In such a framework, the problem (2.2.2) amounts to solve the empirical risk minimization problem

$$\min_{(\mathrm{w},w_0)\in\mathbb{R}^d\times\mathbb{R}} \left\{ \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \mathrm{w}^\top\mathrm{x}_k - w_0\right)^2 \right\}. \tag{2.2.3}$$

Under the assumptions $q < n$ and $\mathrm{X}^\top\mathrm{X}$ is invertible, the solution of problem (2.2.3) is given by

$$\hat{\omega} = (\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top\mathrm{y}.$$

Otherwise, we need to apply some regularization techniques which will be discussed in the next sections.

2.2.2 Neural network

An artificial Neural Network (NN) is a collection of neurons connected together to predict the output y that can be used for regression or classification prediction. The output of one neuron is a function of the weighted input features $\varphi(w_0 + \sum_{k=1}^{d} w_k\mathrm{x}_k)$, where φ is called the activation function and $\omega = (\mathrm{w}, w_0) \in \mathbb{R}^d \times \mathbb{R}$ is called the weight vector with w_0 the intercept.

The activation function quantifies the activation status of the neuron. Several activation functions can be considered

- Heaviside step function,

$$\forall x \in \mathbb{R},\ \varphi(x) = \begin{cases} 1 & \text{if } x \geqslant 0, \\ 0 & \text{otherwise,} \end{cases}$$

- Rectified linear unit (ReLU),

$$\forall x \in \mathbb{R},\ \varphi(x) = \max\{0, x\},$$

- Logisitc function,

$$\forall x \in \mathbb{R},\ \varphi(x) = \frac{1}{1 + e^{-x}},$$

- Hyperbolic tangent,

$$\forall x \in \mathbb{R},\ \varphi(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}.$$

There is no universal rule to select an activation function and such a choice depends on the use cases. Sigmoid functions like logistic or hyperbolic tangent are often used since they lead to good approximation properties as shown in Barron (1993). More generally, this is known that the regularity of the activation function influences the rate of approximation (see Mhaskar and Micchelli (1993) and Kamruzzaman and Aziz (2002)). However, smoothness is not the only factor to take into account and there are many other considerations such as speed of evaluation, difficulty to compute the derivative, preservation of data normalization, ... For some of these reasons, the ReLU activation function is often considered. We can also consider the option of mixing several functions as discussed in Hagg et al. (2017).

The multi-layer perceptron is a neural network proposed by Rosenblatt (1958). It contains multiple layers of neurons. A layer is a set of neurons and each neuron of a layer has a link with the next layer but has no link inside the layer. All outputs of the neurons of this layer are the inputs of the neurons of the next layer. The Figure 2.2-1 shows an architecture of multi-layer perceptron with 3 input variables, one output variable and two hidden layers. Each layer has and intercept neuron associated with its weight w_0 the neuron "1" in Figure 2.2-1.

The multi-layer perceptron has an input layer, hidden layers and output layer. In the input layer, each input variable is represented by a neuron. For the output layer the number of output neurons is directly related to the type of response, for continuous output we have only one output neuron with a specific activation function. The hidden layer is the intermediate layer between input and output layer, the number of hidden layers and neurons are fixed before the estimation of the weights. The more hidden layers we have the larger the number of parameters to estimate which require a regularization to avoid overfitting.

For the multi-layer perceptron, the activation function is applied to the hidden layer and to the output layer neurons for classification prediction. For regression model, the output neuron provides a linear combination of its inputs, which is equivalent to have an identity activation function.

The Gradient Descent (GD) method is an optimization algorithm used to solve the problem (2.2.2) for the NN procedure. GD algorithm is an iterative minimization that goes in the direction of the negative gradient of $\widehat{L}(f_\omega)$. The GD is time consuming and due to the non-linearity

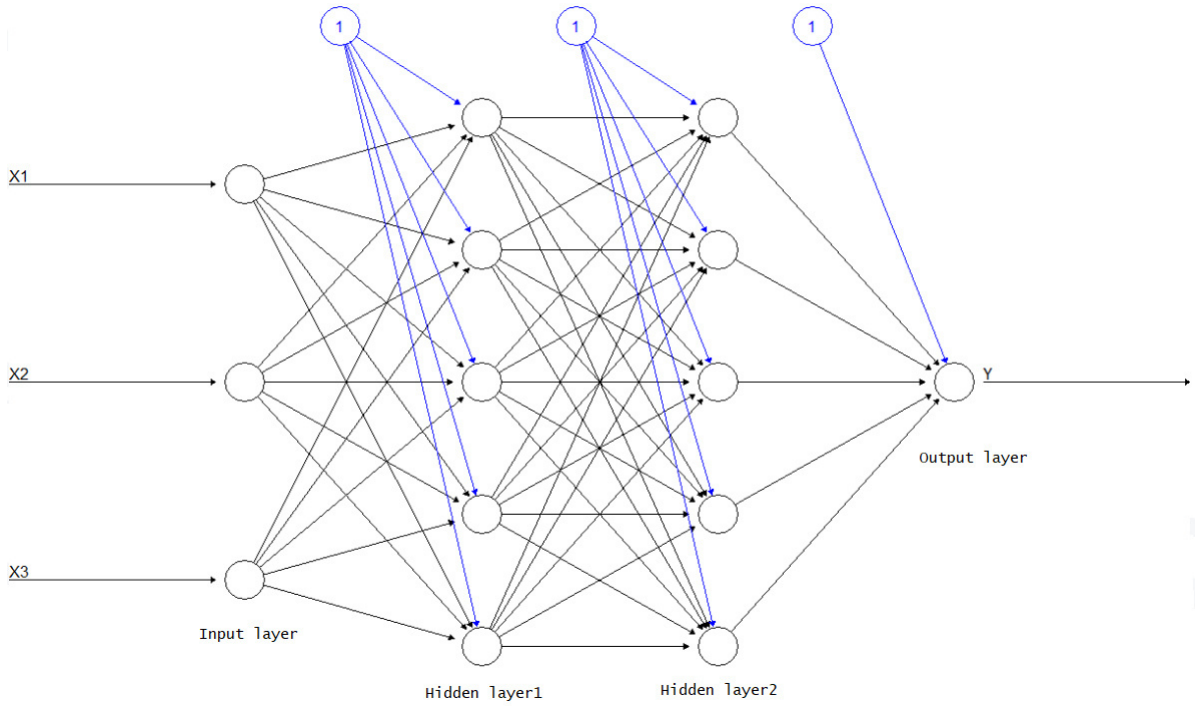

Figure 2.2-1 – Example of multi-layer perceptron with two hidden layers and one output.

introduced by the activation function the GD may get stuck in a local minimum. Some modified versions of GD improve the optimization in terms of speed like the stochastic gradient descent, mini-batch gradient descent, gradient descent with momentum and adaptive moment estimation (Adam). The backpropagation algorithm takes advantage of the perceptron structure to efficiently compute the gradient in order to speed the GD methods.

In practice, we need to define an architecture (number of layers and the number of neurons by layer), the activation function, the optimization method and the number of complete passes through the NN (epochs). There are additional hyperparameters related to the optimization method.

2.2.3 Regression trees based on CART approach

A regression tree is a machine learning method which provides a prediction model based on a binary tree built from training data. The Classification And Regression Tree (CART) model is obtained by recursively splitting the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be represented by a binary tree.

The first step to predict y thanks to the input variables X using CART approach is to define for each variable $j = 1, \ldots, d$ a binary partition of the index set $\{1, \ldots, n\}$ and the split point

$s \in \mathbb{R}$ such that

$$I_1(j,s) = \{i \in \{1,\dots,n\} \text{ such that } x_{ij} \leq s\},$$
$$I_2(j,s) = \{i \in \{1,\dots,n\} \text{ such that } x_{ij} > s\}.$$

Then, we look for the pair (j^*, s^*) that minimizes the loss function :

$$(j^*, s^*) \in \underset{(j,s)}{\arg\min} \left\{ \frac{1}{|I_1(j,s)|} \sum_{i \in I_1(j,s)} (y_i - \bar{y}_1(j,s))^2 + \frac{1}{|I_2(j,s)|} \sum_{i \in I_2(j,s)} (y_i - \bar{y}_2(j,s))^2 \right\}$$

where

$$\bar{y}_1(j,s) = \frac{1}{|I_1(j,s)|} \sum_{i \in I_1(j,s)} y_i, \quad \bar{y}_2(j,s) = \frac{1}{|I_2(j,s)|} \sum_{i \in I_2(j,s)} y_i$$

and $|.|$ represents the cardinality of a set.

This quantity corresponds to the loss function of a piecewise constant function with fixed value on each subspace $I_1(j^*, s^*)$ and $I_2(j^*, s^*)$. On the resulting subsets I_1 and I_2, we apply the same procedure and iterate until the subsets contain only one observation or a stopping rule is satisfied (variance criterion, depth limit, ...).

The CART procedure is known to be unstable and the bagging procedure gives some stability by adding randomness to the CART which reduces the prediction error. The idea of the bagging procedure is to draw B new training sets with replacement from X and y denoted by X_b and y_b with $b = 1, \dots, B$. Theses samples are known as a bootstrap samples.

A decision tree is built for each bootstrap b. The output variable is fitted by averaging the predictions from all the individual regression trees.

The random-forest regression add randomness to the bagging procedure during the learning process. For each bootstrap, at each splitting iteration, the decision tree is built by selecting m explanatory variables draw randomly from d variables .

2.3 Dropout regularization

In this section we study the dropout regularization applied to the selected algorithms to solve the minimization problem (2.2.2). From a statistical point of view, any minimizer $\hat{\omega}$ which is a solution to this problem leads to a predictor $f_{\hat{\omega}}$ trained to fit the data.

Assuming $\widehat{L}(f_\omega)$ is regular enough with respect to ω, a naive algorithm to solve the problem (2.2.2) consists in considering a gradient descent on the objective function $\omega \mapsto \widehat{L}(f_\omega)$. Thus, we

build a sequence $(\omega_t)_{t\geqslant 0}$ whose iterates are given by the recursion

$$\forall t \geqslant 0,\ \omega_{t+1} = \omega_t - \eta \nabla \widehat{L}(f_{\omega_t}) \tag{2.3.1}$$

where $\omega_0 \in \Omega$ is an initial state and $\eta > 0$ is a fixed learning rate. Assuming further that the objective function is convex, we know that $\widehat{L}(f_{\omega_t})$ converges towards a solution to the problem (2.2.2) as soon as the learning rate is sufficiently small compared to the inverse of the Lipschitz constant of the gradient (see Nesterov (2013)). In practice, such a procedure can be hard to use because of the time required for an explicit computation of the gradient $\nabla \widehat{L}(f_{\omega_t})$ at each iteration based on the whole data set. The method may even become unusable when the data set is too large and some stochastic approaches described later can be implemented in order to circumvent this limitation.

A solution to the minimization problem (2.2.2) generally leads to a good accuracy on the initial data, called training data set. This phenomenon is particularly evident when the parameter dimension q is large with respect to the number of observations n. The well-known drawback is then a poor quality of the predictions based on new data, called testing data set. Such a phenomenon is named overfitting and it is advisable to limit it with regularization methods by adding a penalization term to the empirical risk. There are many ways to regularize, we are particularly interested here in the dropout approach. The first work that introduce the dropout technique was done by Hinton et al. (2012) where they demonstrated the effectiveness of dropout on the test error. Since the victory of Srivastava et al. (2014) at the ImageNet Large Scale Visual Recognition Challenge, the dropout technique generate many of interest and become the regularization the most used in neural networks since it is known to provide efficient generalization capacities of overparameterized models.

In this section, we give an explicit expression of dropout regularization for linear model and single layer perceptron. We give an experimental comparison between two ways of dropout for a single layer perceptron. We also add an experimental results for the dropout applied to neural network and CART procedure.

2.3.1 Dropout regularization for linear regression

In order to do not apply any assumption on the empirical covariance matrix associated to the vectors $\mathrm{x}_1, \ldots, \mathrm{x}_n$, we use a gradient descent minimizer (see Equation (2.3.1)) to solve the problem (2.2.3) as detailed in Algorithm 1 for a fixed number of epochs, i.e. a fixed number of times the algorithm walks through the whole data set.

For linear regression, dropout amounts to regularize the minimization problem (2.2.3) by keeping only some input features uniformly at random at training time. As described in Mianjy et al.

Algorithm 1: Gradient Descent for Linear Regression

Data: $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)$
Input: Number of epochs T, learning rate η
Initialize $\omega_0 = (\mathrm{w}_0, w_{0,0})$
for $t = 0, \ldots, T-1$ **do**

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \frac{2\eta}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}_t^\top \mathrm{x}_k - w_{0,t}\right) \mathrm{x}_k$$

$$w_{0,t+1} \leftarrow (1 - 2\eta) w_{0,t} + \frac{2\eta}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}_t^\top \mathrm{x}_k\right)$$

$$\omega_{t+1} \leftarrow (\mathrm{w}_{t+1}, w_{0,t+1})$$

end
Output: ω_T

(2018), this procedure can be seen as a stochastic gradient descent where the input features are randomly dropped at each iteration, which consists in considering the objective function

$$\omega = (\mathrm{w}, w_0) \mapsto \widehat{L}_{\mathrm{drop}}(f_\omega) = \mathbb{E}_{\mathrm{b}} \left[\frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \operatorname{diag}(\mathrm{b}) \mathrm{x}_k - w_0\right)^2 \right]$$

where b is a vector whose entries are i.i.d. random variables distributed as $(1-p)^{-1}\mathrm{Ber}(1-p)$. The parameter $p \in [0, 1)$ is therefore the probability of dropping an input feature. Note that the dropout does not affect the intercept w_0 in the definition above. Such a choice is common in practice since it makes the procedure independent from the origin chosen for output y (see Chapter 3 of Hastie et al. (2009)). Moreover, this also allows us to extend the dropout technique to neural networks in the following sections while keeping similar notations.

As a stochastic gradient descent, dropout provides a sequence of random parameters $(\omega_t)_{t \geqslant 0}$ defined by the recursion

$$\forall t \geqslant 0, \ \omega_{t+1} = \omega_t - \eta \nabla \widehat{L}_{\mathrm{drop}}^{(t+1)}(f_{\omega_t})$$

where $\omega_0 \in \mathbb{R}^d \times \mathbb{R}$ is an initial state, $\eta > 0$ is a fixed learning rate and the random functions $\omega \mapsto \widehat{L}_{\mathrm{drop}}^{(t+1)}(f_\omega)$ are given by

$$\omega = (\mathrm{w}, w_0) \mapsto \widehat{L}_{\mathrm{drop}}^{(t+1)}(f_\omega) = \frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \operatorname{diag}(\mathrm{b}_{t+1}) \mathrm{x}_k - w_0\right)^2.$$

The dropping vectors $(\mathrm{b}_t)_{t \geqslant 1}$ are i.i.d. random vectors distributed as b and sample which input features are concerned by the current gradient descent step. Moreover, the obtained iterates are

often averaged into the sequence of random parameters $(\bar{\omega}_t)_{t\geqslant 0}$ where $\bar{\omega}_0 = \omega_0$ and

$$\forall t \geqslant 0,\ \bar{\omega}_{t+1} = \frac{1}{t+1}\sum_{k=1}^{t+1}\omega_k = \bar{\omega}_t + \frac{\omega_{t+1} - \bar{\omega}_t}{t+1},$$

since it is well-known that the convergence behavior is then better (see Polyak and Juditsky (1992) and Ruppert (1988)). Algorithm 2 summarizes the main steps of this stochastic optimization procedure, which nevertheless still requires an explicit computation of a gradient at each iteration.

Algorithm 2: Dropout with Gradient Descent for Linear Regression

Data: $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)$
Input: Number of epochs T, learning rate η, dropout rate p
Initialize $\omega_0 = \bar{\omega}_0 = (\mathrm{w}_0, w_{0,0})$
for $t = 0, \ldots, T-1$ **do**
  Sample vector b_{t+1} entries as i.i.d. $(1-p)^{-1}\mathrm{Ber}(1-p)$
  $\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \dfrac{2\eta}{n}\displaystyle\sum_{k=1}^{n}\left(y_k - \mathrm{w}_t^\top \operatorname{diag}(\mathrm{b}_{t+1})\mathrm{x}_k - w_{0,t}\right)\operatorname{diag}(\mathrm{b}_{t+1})\mathrm{x}_k$
  $w_{0,t+1} \leftarrow (1-2\eta)w_{0,t} + \dfrac{2\eta}{n}\displaystyle\sum_{k=1}^{n}\left(y_k - \mathrm{w}_t^\top \operatorname{diag}(\mathrm{b}_{t+1})\mathrm{x}_k\right)$
  $\omega_{t+1} \leftarrow (\mathrm{w}_{t+1}, w_{0,t+1})$
  $\bar{\omega}_{t+1} \leftarrow \bar{\omega}_t + (\omega_{t+1} - \bar{\omega}_t)/(t+1)$
end
Output: ω_T and $\bar{\omega}_T$

The connection between dropout and the regularization of the problem (2.2.3) appears through a straightforward calculation (see details in Appendix 4.5), for any $\omega = (\mathrm{w}, w_0) \in \mathbb{R}^d \times \mathbb{R}$,

$$\widehat{L}_{\mathrm{drop}}(f_\omega) = \widehat{L}(f_\omega) + \frac{p}{1-p}\|\Gamma \mathrm{w}\|^2 \tag{2.3.2}$$

where $\Gamma^2 = \operatorname{ddiag}\left(\frac{1}{n}\sum_{k=1}^{n}\mathrm{x}_k\mathrm{x}_k^\top\right)$.

This equation illustrates how dropout acts as a regularizer by adding an additional term to the function $\widehat{L}(f_\omega)$ to be minimized. The regularization term $\|\Gamma \mathrm{w}\|^2$ is known as Tikhonov regularizer and the dropout rate p only appears in the multiplicative factor $p(1-p)^{-1}$. Since the intercept w_0 is left out the dropout, this coefficient is not penalized. When Γ is the identity matrix, e.g. for normalised data, we recognize a common ridge penalty term but, in a general setting, the penalty is data-dependent. Such a link between feature corruption and regularization is well-known. For example, it has been discussed in Bishop (1995) for additive Gaussian noise, in Heinze et al. (2014) for random Gaussian projection or in Wang and Manning (2013) and Wager et al. (2013) for a dropout approach similar to ours.

As indicated above, the dropout implementation given in Algorithm 2 is based on descent steps computed on the whole data set with gradients of randomly perturbed empirical means. The randomness of the gradient descent only concerns the dropout of the input features at each iteration. If the size of the data set is large, such an explicit gradient is still a problem and the method is limited in practice. To get around this limitation, a stochastic approach can also be used for the empirical mean seen as the expectation according to the uniform distribution on the data set. Such an approach is common in machine learning and can be implemented by splitting the training data set into random subsets, called batches, whose size is small enough to allow a gradient calculation used as an approximation for the current descent step. Moreover, it is known that the learning rate has to decay together with averaging to obtain fast convergence rates in such a stochastic approach (see Bach and Moulines (2011)). Thus, the learning rate is no longer fixed but defined by a sequence $(\eta_t)_{t\geqslant 1}$ of positive real numbers to construct a sequence of random parameters $(\omega_t)_{t\geqslant 0}$ based on the recursion

$$\forall t \geqslant 0,\ \omega_{t+1} = \omega_t - \eta_{t+1} \nabla_{B_{t+1}} \widehat{L}^{(t+1)}_{\text{drop}}(f_{\omega_t})$$

where $\omega_0 \in \mathbb{R}^d \times \mathbb{R}$ is an initial state, $(B_t)_{t\geqslant 1}$ is a sequence of random batches and ∇_B stands for the gradient operator restricted to batch $B \subset \{1, \ldots, n\}$. In practice, a polynomial decay for the learning rate is often considered. This is also advisable to sample the batches involved in a given epoch as a partition of the training data set, even if the disjoint subsets do not all have exactly the same size. This procedure detailed in Algorithm 3 is the one we use for linear regression in the section devoted to the experimental study.

Algorithm 3: Dropout with Stochastic Gradient Descent for Linear Regression

Data: $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)$
Input: Number of epochs T, number of batches per epoch M, learning rates $(\eta_t)_{t\geqslant 1}$, dropout rate p
Initialize $\omega_0 = \bar{\omega}_0 = (\mathrm{w}_0, w_{0,0})$
for $t' = 0, \ldots, T-1$ **do**
  Sample random disjoint batches $B_{t'M+1}, \ldots, B_{(t'+1)M}$
  for $t = t'M, \ldots, (t'+1)M - 1$ **do**
    Sample vector b_{t+1} entries as i.i.d. $(1-p)^{-1}\mathrm{Ber}(1-p)$
    $\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \dfrac{2\eta_{t+1}}{|B_{t+1}|} \displaystyle\sum_{k \in B_{t+1}} \left(y_k - \mathrm{w}_t^\top \operatorname{diag}(\mathrm{b}_{t+1})\mathrm{x}_k - w_{0,t}\right) \operatorname{diag}(\mathrm{b}_{t+1})\mathrm{x}_k$
    $w_{0,t+1} \leftarrow (1 - 2\eta_{t+1}) w_{0,t} + \dfrac{2\eta_{t+1}}{|B_{t+1}|} \displaystyle\sum_{k \in B_{t+1}} \left(y_k - \mathrm{w}_t^\top \operatorname{diag}(\mathrm{b}_{t+1})\mathrm{x}_k\right)$
    $\omega_{t+1} \leftarrow (\mathrm{w}_{t+1}, w_{0,t+1})$
    $\bar{\omega}_{t+1} \leftarrow \bar{\omega}_t + (\omega_{t+1} - \bar{\omega}_t)/(t+1)$
  end
end
Output: ω_{TM} and $\bar{\omega}_{TM}$

2.3.2 Dropout regularization with single layer perceptron

Let us now focus on simple feedforward neural networks aggregating the outputs of a single hidden layer. The core of such a predictor is a single layer perceptron as invented by Rosenblatt (1958). To this end, we consider an activation function $\varphi : \mathbb{R} \to \mathbb{R}$, a weight matrix $\mathrm{V}^{(1)} \in \mathbb{R}^{d_1 \times d}$ and an intercept vector $\mathrm{v}_0^{(1)} \in \mathbb{R}^{d_1}$ to define the perceptron output vector $\varphi(\mathrm{V}^{(1)}\mathrm{x} + \mathrm{v}_0^{(1)}) \in \mathbb{R}^{d_1}$ associated to an input vector $\mathrm{x} \in \mathbb{R}^d$. Being given $\mathrm{w} \in \mathbb{R}^{d_1}$ and $w_0 \in \mathbb{R}$, we can then introduce the predictor function f_ω obtained by a linear combination of the perceptron outputs,

$$\forall \mathrm{x} \in \mathbb{R}^d, \ f_\omega(\mathrm{x}) = \mathrm{w}^\top \varphi\left(\mathrm{V}^{(1)}\mathrm{x} + \mathrm{v}_0^{(1)}\right) + w_0$$

where we have set the parameter $\omega = (\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}, \mathrm{w}, w_0) \in \mathbb{R}^{d_1 \times d} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \times \mathbb{R}$. The predictor class of such functions leads to solve the empirical risk minimization problem

$$\min_{(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}, \mathrm{w}, w_0) \in \mathbb{R}^{d_1 \times d} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \times \mathbb{R}} \left\{ \frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \varphi\left(\mathrm{V}^{(1)}\mathrm{x}_k + \mathrm{v}_0^{(1)}\right) - w_0 \right)^2 \right\}. \tag{2.3.3}$$

The neural network is said to have 2 layers because of the stacking of matrix products and the activation function. The middle layer of neurons is known as a hidden layer of size d_1.

In practice, the activation function φ defines the behavior of a neuron output with respect to the linear combination of inputs around an intercept value. Common examples of such a function include : Heaviside step, Logisitc Hyperbolic tangent or ReLU. In this work, we consider neural networks with the same activation function for all neurons.

The aim of this section is to use dropout to regularize the minimization problem (2.3.3). A first way consists in dropping only in the linear combination of the perceptron outputs. As discussed in Section 2.3.1, dropout is seen as a stochastic gradient descent and we thus consider the objective function

$$\omega \mapsto \widehat{L}_{\mathrm{drop}}(f_\omega) = \mathbb{E}_{\mathrm{b}} \left[\frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \operatorname{diag}(\mathrm{b}) \varphi\left(\mathrm{V}^{(1)}\mathrm{x}_k + \mathrm{v}_0^{(1)}\right) - w_0 \right)^2 \right]$$

where b is a vector whose entries are i.i.d. random variables distributed as $(1-p)^{-1}\mathrm{Ber}(1-p)$. The implementation of this method that we use in the sequel is based on a mini batch stochastic gradient descent as presented in Algorithm 4.

Arguing the same way we did for linear regression (see details in Appendix 4.5), the regularization of the problem (2.3.3) is given by the following relation, for any $\omega = (\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}, \mathrm{w}, w_0) \in \mathbb{R}^{d_1 \times d} \times \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \times \mathbb{R}$,

$$\widehat{L}_{\mathrm{drop}}(f_\omega) = \widehat{L}(f_\omega) + \frac{p}{1-p} \left\| \Gamma\left(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}\right) \mathrm{w} \right\|^2 \tag{2.3.4}$$

Algorithm 4: Dropout with Single Layer Perceptron (differentiable activation function)

Data: $(\mathrm{x}_1, y_1), \ldots, (\mathrm{x}_n, y_n)$

Input: Number of epochs T, number of batches per epoch M, learning rates $(\eta_t)_{t\geqslant 1}$, dropout rate p

Initialize $\omega_0 = \bar{\omega}_0 = (\mathrm{V}_0^{(1)}, \mathrm{v}_{0,0}^{(1)}, \mathrm{w}_0, w_{0,0})$

for $t' = 0, \ldots, T-1$ **do**

Sample random disjoint batches $B_{t'M+1}, \ldots, B_{(t'+1)M}$

for $t = t'M, \ldots, (t'+1)M - 1$ **do**

Sample vector b_{t+1} entries as i.i.d. $(1-p)^{-1}\mathrm{Ber}(1-p)$

$$z_{k,t} \triangleq y_k - \mathrm{w}_t^\top \operatorname{diag}(\mathrm{b}_{t+1})\varphi\left(\mathrm{V}_t^{(1)}\mathrm{x}_k + \mathrm{v}_{0,t}^{(1)}\right) - w_{0,t}$$

$$\mathrm{V}_{t+1}^{(1)} \leftarrow \mathrm{V}_t^{(1)} + \frac{2\eta_{t+1}}{|B_{t+1}|}\sum_{k\in B_{t+1}} z_{k,t}\left(\mathrm{w}\odot\operatorname{diag}(\mathrm{b}_{t+1})\varphi'\left(\mathrm{V}_t^{(1)}\mathrm{x}_k + \mathrm{v}_{0,t}^{(1)}\right)\right)\mathrm{x}_k^\top$$

$$\mathrm{v}_{0,t+1}^{(1)} \leftarrow \mathrm{v}_{0,t}^{(1)} + \frac{2\eta_{t+1}}{|B_{t+1}|}\sum_{k\in B_{t+1}} z_{k,t}\left(\mathrm{w}\odot\operatorname{diag}(\mathrm{b}_{t+1})\varphi'\left(\mathrm{V}_t^{(1)}\mathrm{x}_k + \mathrm{v}_{0,t}^{(1)}\right)\right)$$

$$\mathrm{w}_{t+1} \leftarrow \mathrm{w}_t + \frac{2\eta_{t+1}}{|B_{t+1}|}\sum_{k\in B_{t+1}} z_{k,t}\operatorname{diag}(\mathrm{b}_{t+1})\varphi\left(\mathrm{V}_t^{(1)}\mathrm{x}_k + \mathrm{v}_{0,t}^{(1)}\right)$$

$$w_{0,t+1} \leftarrow (1-2\eta_{t+1})w_{0,t} + \frac{2\eta_{t+1}}{|B_{t+1}|}\sum_{k\in B_{t+1}}\left(y_k - \mathrm{w}_t^\top\operatorname{diag}(\mathrm{b}_{t+1})\varphi\left(\mathrm{V}_t^{(1)}\mathrm{x}_k + \mathrm{v}_{0,t}^{(1)}\right)\right)$$

$$\omega_{t+1} \leftarrow (\mathrm{w}_{t+1}, w_{0,t+1})$$

$$\bar{\omega}_{t+1} \leftarrow \bar{\omega}_t + (\omega_{t+1} - \bar{\omega}_t)/(t+1)$$

end

end

Output: ω_{TM} and $\bar{\omega}_{TM}$

where

$$\Gamma^2\left(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}\right) = \mathrm{ddiag}\left(\frac{1}{n}\sum_{k=1}^{n}\varphi\left(\mathrm{V}^{(1)}\mathrm{x}_k + \mathrm{v}_0^{(1)}\right)\varphi\left(\mathrm{V}^{(1)}\mathrm{x}_k + \mathrm{v}_0^{(1)}\right)^\top\right).$$

The Tikhonov regularizer is then related to the empirical second moment of the output of the hidden neurons. Such a data-dependent regularization is difficult to study in this general form. Nevertheless, we know that capacity control is thus obtained in the neural network since the squared norm is similar to a l_2-path-norm regularizer as studied in Neyshabur et al. (2015). Assuming the input distribution D is symmetric and isotropic (e.g. standard Gaussian distribution), the authors of Mianjy et al. (2018) zero the intercept vector $\mathrm{v}_0^{(1)}$ and show that, if we consider the ReLU activation function, the expected regularizer is exactly given by the l_2-path-norm of the neural network, namely

$$\mathbb{E}_D\left[\left\|\Gamma\left(\mathrm{V}^{(1)}, 0\right)\mathrm{w}\right\|^2\right] = \frac{1}{2}\sum_{i=1}^{d_1}\sum_{j=1}^{d}\mathrm{w}_i^2 {V_{ij}^{(1)}}^2.$$

Dropout can also affect the input layer by dropping some variables at random. To this end, another random vector is introduced in the expectation to define the objective function

$$\omega \mapsto \widehat{L}'_{\mathrm{drop}}(f_\omega) = \mathbb{E}_{\mathrm{b},\mathrm{b}^{(1)}}\left[\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \mathrm{w}^\top \mathrm{diag}(\mathrm{b})\varphi\left(\mathrm{V}^{(1)}\,\mathrm{diag}(\mathrm{b}^{(1)})\mathrm{x}_k + \mathrm{v}_0^{(1)}\right) - w_0\right)^2\right]$$

where b and $\mathrm{b}^{(1)}$ are independent vectors whose entries are i.i.d. random variables distributed as $(1-p)^{-1}\mathrm{Ber}(1-p)$ and $(1-p^{(1)})^{-1}\mathrm{Ber}(1-p^{(1)})$, respectively, with $p, p^{(1)} \in [0,1)$. This technique can be implemented in a variant of Algorithm 4 by sampling $\mathrm{b}_{t+1}^{(1)}$ at each step t and multiplying the input vectors x_k by $\mathrm{diag}(\mathrm{b}_{t+1}^{(1)})$.

2.3.3 Dropout regularization for neural networks

Dropout is widely used in practice to train neural networks while avoiding overfitting and we focus here on shallow feedforward neural networks. Namely, we consider stackings of linear combination and activation function $\varphi : \mathbb{R} \to \mathbb{R}$ as introduced in the previous section. For the sake of readability, if $(\mathrm{V}, \mathrm{v}_0) \in \mathbb{R}^{d'\times d} \times \mathbb{R}^{d'}$ are the parameters of a hidden layer, we denote $\varphi_{\mathrm{V},\mathrm{v}_0}$ the associated function

$$\forall \mathrm{x} \in \mathbb{R}^d,\ \varphi_{\mathrm{V},\mathrm{v}_0}(\mathrm{x}) = \varphi\left(\mathrm{V}\mathrm{x} + \mathrm{v}_0\right) \in \mathbb{R}^{d'}.$$

Let $d_0 = d$ and $K \geqslant 1$ be the number of hidden layers of respective sizes $d_1, \ldots, d_K > 0$, the predictor class is then given by functions with the form

$$\forall \mathrm{x} \in \mathbb{R}^d,\ f_\omega(\mathrm{x}) = \mathrm{w}^\top \varphi_{\mathrm{V}^{(K)},\mathrm{v}_0^{(K)}} \circ \cdots \circ \varphi_{\mathrm{V}^{(1)},\mathrm{v}_0^{(1)}}(\mathrm{x}) + w_0$$

where, for any $j \in \{1, \dots, K\}$, $(V^{(j)}, \mathrm{v}_0^{(j)}) \in \mathbb{R}^{d_j \times d_{j-1}} \times \mathbb{R}^{d_j}$ and $(\mathrm{w}, w_0) \in \mathbb{R}^{d_K} \times \mathbb{R}$. Hereafter, we denote by $\omega = ((V^{(1)}, \mathrm{v}_0^{(1)}), \dots, (V^{(K)}, \mathrm{v}_0^{(K)}), (\mathrm{w}, w_0))$ the parameters of such a function. Thus, the empirical risk minimization problem to solve here can be written as

$$\min_{\substack{\forall j,\ (V^{(j)}, \mathrm{v}_0^{(j)}) \in \mathbb{R}^{d_j \times d_{j-1}} \times \mathbb{R}^{d_j} \\ (\mathrm{w}, w_0) \in \mathbb{R}^{d_K} \times \mathbb{R}}} \left\{ \frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \varphi_{\mathrm{V}^{(K)}, \mathrm{v}_0^{(K)}} \circ \cdots \circ \varphi_{\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}}(\mathrm{x}_k) - w_0 \right)^2 \right\}. \tag{2.3.5}$$

Note that the number of parameters to be tuned is equal to $d_1(d_0+1)+\cdots+d_K(d_{K-1}+1)+d_K+1$ and such a model can quickly become overparameterized when the number of layers increase. In this case, regularization is not optional and has to be implemented to ensure good behavior of the neural network.

As discussed in the previous section, dropout can independently affect each layer of the neural network, be it input or hidden ones. The objective function $\omega \mapsto \widehat{L}_{\mathrm{drop}}(f_\omega)$ involved in the stochastic gradient descent is then given by

$$\mathbb{E}_{\mathrm{b}, \mathrm{b}^{(1)}, \dots, \mathrm{b}^{(K)}} \left[\frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \operatorname{diag}(\mathrm{b}) \varphi_{\mathrm{V}^{(K)} \operatorname{diag}(\mathrm{b}^{(K)}), \mathrm{v}_0^{(K)}} \circ \cdots \circ \varphi_{\mathrm{V}^{(1)} \operatorname{diag}(\mathrm{b}^{(1)}), \mathrm{v}_0^{(1)}}(\mathrm{x}_k) - w_0 \right)^2 \right]$$

where b and $\mathrm{b}^{(j)}$, $j \in \{1, \dots, K\}$, are independent vectors whose entries are i.i.d. random variables distributed as $(1-p)^{-1}\mathrm{Ber}(1-p)$ and $(1-p^{(j)})^{-1}\mathrm{Ber}(1-p^{(j)})$, respectively, with $p, p^{(j)} \in [0, 1)$ and $j \in \{1, \dots, K\}$. To implement this technique, an explicit calculation of the gradient of $\widehat{L}_{\mathrm{drop}}(f_\omega)$ is still straightforward using the Hadamard product. Such an approach is known as forward propagation since it represents how the initial information provided by the input x propagates through the hidden layers to produce the output y. However, the numerical evaluation of the gradient can be computationally expensive due to the number of parameters. This drawback leads to inefficient implementation and therefore we do not provide a detailed algorithm here.

To efficiently compute the gradient of the objective function for a feedforward neural network, it is convenient to use the structure given by the fully connected consecutive layers. Indeed, as discussed in Chapter 2 of Nielsen (2015) and in Chapter 6 of Goodfellow et al. (2016), the composition of functions makes it possible to use a fast matrix method to efficiently evaluate this gradient starting from the loss function and flowing backward through the neural network towards the input layer. Such a way to compute a gradient is known as backpropagation and was originally introduced in Rumelhart et al. (1986). It is beyond the scope to give here the details of backpropagation since this algorithm is already well implemented in widely used neural network libraries. Nevertheless, it is fruitful to keep in mind an example given in Goodfellow et al. (2016) to illustrate the propagation phenomenon and to motivate its use in practice. Thus, we consider a function $g : \mathbb{R} \to \mathbb{R}$, some input $u \in \mathbb{R}$ and a composition chain $x = g(u)$, $y = g(x)$ and $z = g(y)$.

To compute the gradient of z with respect to u, we apply the composition rule to obtain

$$\frac{\partial z}{\partial w} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}\frac{\partial x}{\partial w} = g'(g(g(w)))g'(g(w))g'(w).$$

Thinking g as a layer of the neural network shows that evaluating the value $g(u)$ only once and store it to reuse in upper layers is computationally efficient. Doing the same for $g(g(u))$ and so on is the approach taken by the backpropagation algorithm.

Dropout is known to regularize the problem (2.3.5) in practice. Generally, it is advised to first run the procedure with a dropout rate around 0.2 for all the layers. However, this is only a recommended practice because there is no theoretical result to justify it. Indeed, the explicit calculation of the dropout regularizer in this framework is not feasible with the theoretical tools currently available. Moreover, the way to decrease the learning rate of the associated stochastic gradient descent is still a debated problem. We propose to illustrate these points experimentally in the next subsection using the implementation of neural networks and dropout technique provided by the Tensorflow library (see Abadi and *et al.* (2015)).

2.3.4 Dropout regularization and random-forest

The random-forest algorithm acts as a regularization procedure for the CART bagging by adding randomness in the choice of explanatory variables during the splitting iterations. This can be seen as a way of dropout applied to the input variables. The number of variable selected m can be expressed using the dropout rate by $m = \lfloor (1-p)\ d \rfloor$ where $\lfloor . \rfloor$ stands for the floor part function.

2.3.5 Experimental study

To run our experimentation, we consider 200 flights of distinct lengths from a given aircraft. To be able to predict the oil temperature y using the history of the explanatory variables, we split the flights into segments as described in Section 3.2. The explanatory variables selected to this end are given in Table 2.3.1.

For each segment of flight, we expand the observed explanatory variables onto Fourier basis as explained in Section 2.1. For our experimentation we set $K = 31$ coefficients by explanatory variables which leads to a data set of 7442×217 dimensions. We divide the data set into training (80% of data set) and test (20% of data set) sets.

The aim of this experimentation is to understand how the dropout acts on the Mean Square Error (MSE) (see Equation (2.2.1)). For that, we compare MSE computed on the training set and the MSE computed on the test set. To evaluate the dropout effect we use the same approach as Arora et al. (2020), by computing the generalization gap which is the difference between MSE test

Table 2.3.1 – Explanatory variables used to predict the generator oil temperature

Explanatory variable	**Unit**
Generator oil temperature (y)	C°
Engine speed	Knot (kts)
Static air temperature	C°
Total air temperature	C°
Computed air speed	kts
Altitude	ft
Generator load	KVA
Delta international standard atmosphere	-

and the MSE training.

Linear model

For linear model experimentation, we use Algorithm 3 and compare the different dropout rates $p = 0.1$, 0.2, 0.3 and 0.4 between them and with the no regularized one ($p = 0$). We set the weight vector ω_0 to the zero vector, the number of epochs to $T = 30000$, the initial learning rate to $\eta_0 = 10^{-3}$ and the number of batches per epoch to $M = 5$.

In Figure 2.3-2 we represent the progress of the MSE with respect to the number of epochs for the training data set (left), the test data set (middle) and the generalization gap (right). The overfitting phenomena is clearly visible for the curve SGD ($p = 0$) whereas the curves with dropout show a MSE of test set smaller than the MSE of trainning set. The smallest MSE observed in the test set is for a dropout rate $p = 0.2$.

Neural network

For all the experimentations on NN, we use Tensorflow libraries in Python. We chose the activation function ReLU and the sgd optimizer. We set the learning rate to $\eta = 0.001$ with a linear decay. All hidden layers have the same number of units set to $2 \times d$ (434) to reach easily the overfitting of the training set.

For the single layer perceptron (one hidden layer with 434 units), we apply a different value of dropout rate on the hidden layer as described in Equation (2.3.4) and set the number of epochs to $T = 60000$. Figure 2.3-3 shows clearly the regularization benefits on the accuracy of NN. In the training set plot (left), the SGD without dropout gets the best MSE compared to the dropout SGD curves. But as soon as we try to fit a new data only the dropout SGD curves remain efficient (middle). The dropout rates $p \geq 0.2$ have a similar behavior while the dropout rates $p = 0.1$ shows a slight overfitting phenomena.

To study further the role of p on the NN accuracy, in Figure 2.3-4, we compare the MSE of the test set for several NN architectures. We compare architectures with one, two and three hidden

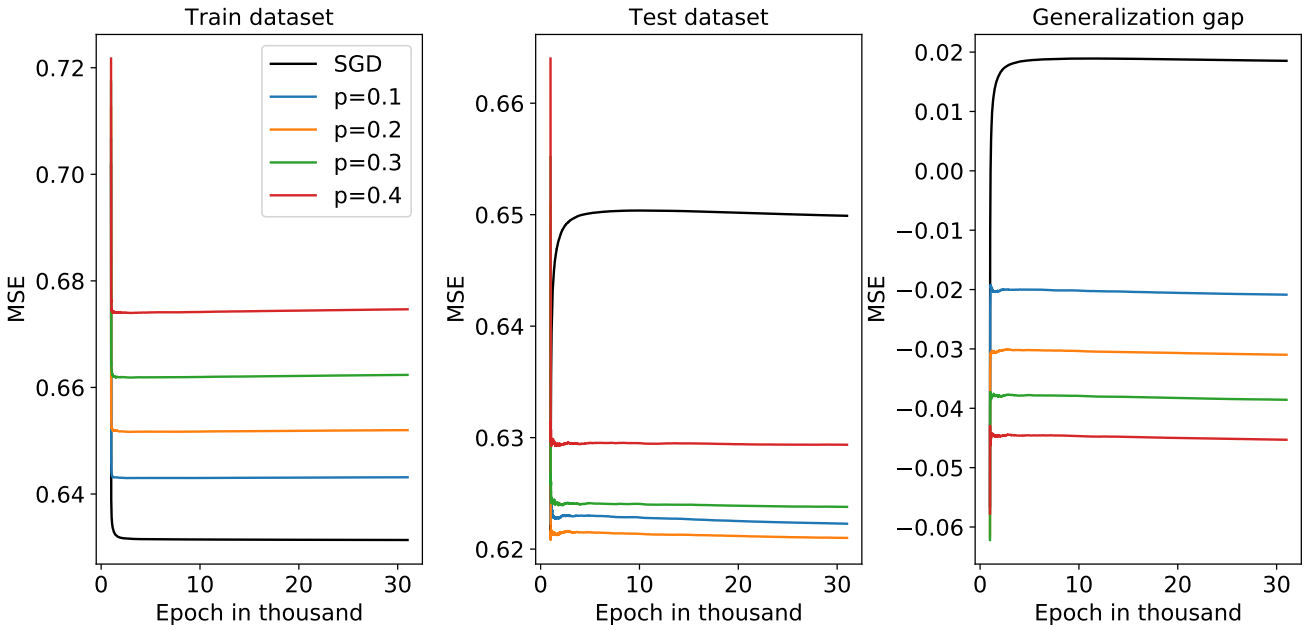

Figure 2.3-2 – Dropout application on a linear model with a dropout rates of $p = 0$ (SGD), 0.1, 0.2, 0.3 and 0.4. Left: MSE of training set. Center: MSE of test set. Right: Difference between the MSE test and the MSE train

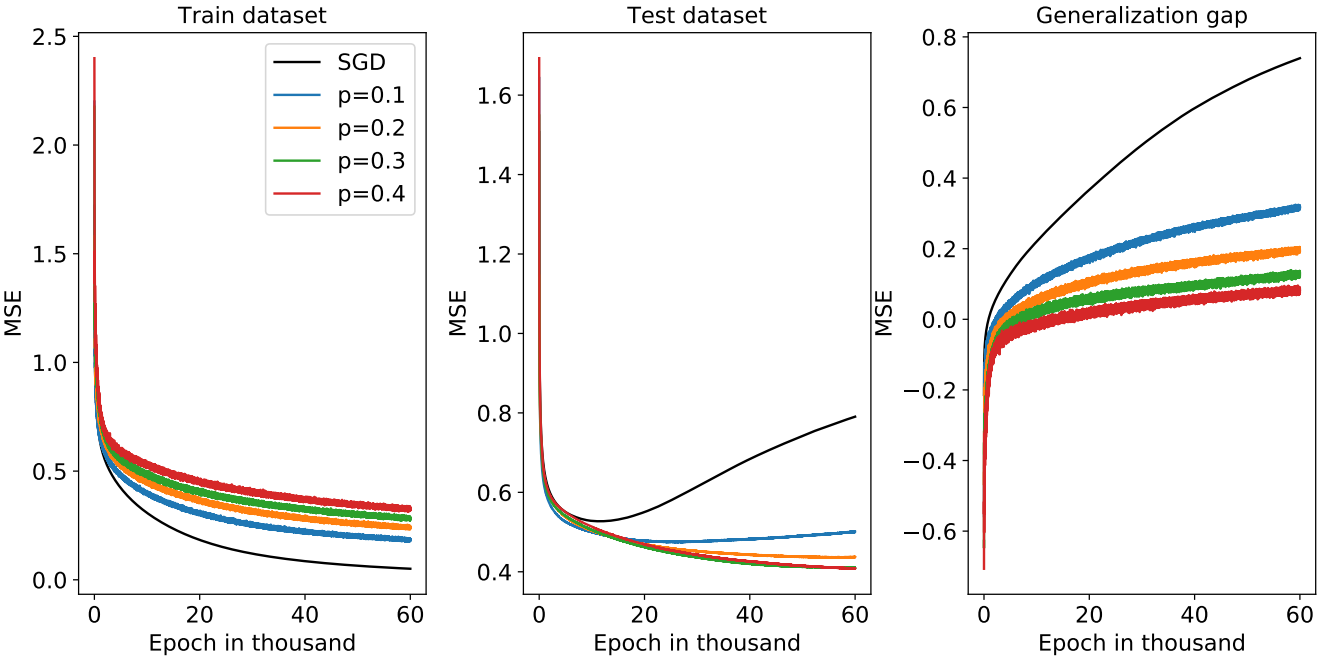

Figure 2.3-3 – Dropout application on a the hidden layer of single layer perceptron with a dropout rate $p = 0$ (SGD), 0.1, 0.2, 0.3 and 0.4. Left: MSE of train set. Center: MSE of test set. Right: Difference between the MSE test and the MSE train.

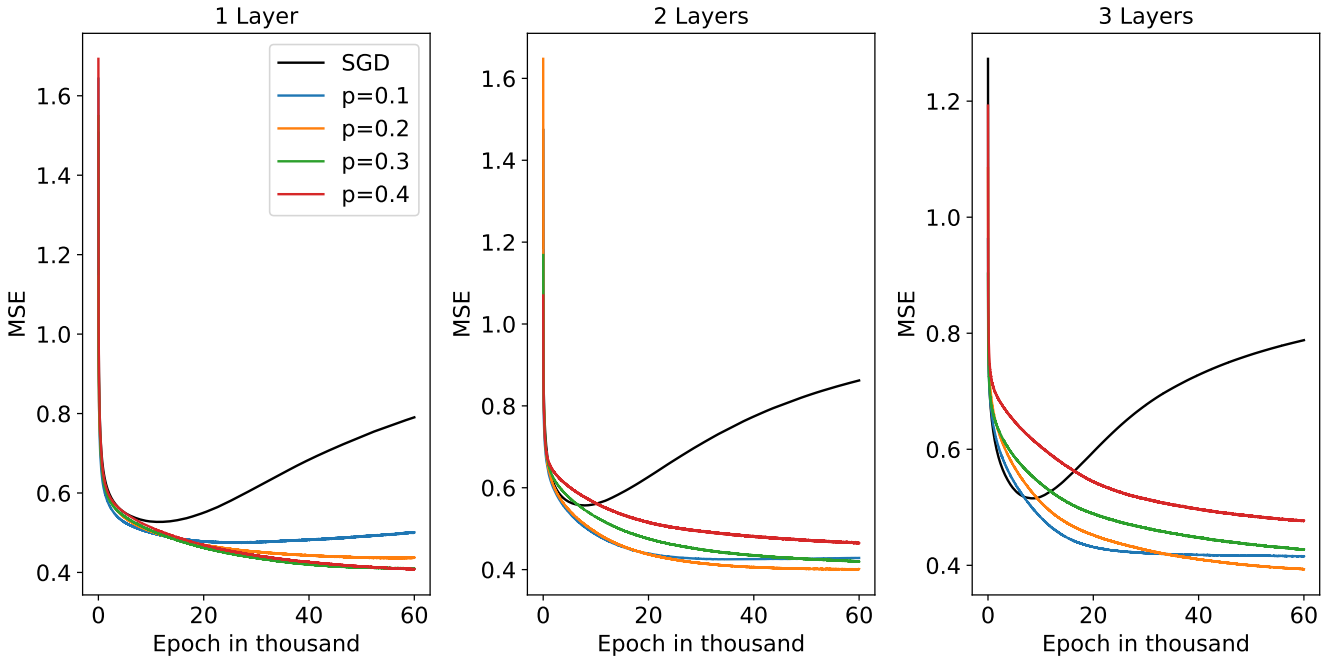

Figure 2.3-4 – MSE of test set of different NN architecture with $p = 0$ (SGD), 0.1, 0.2, 0.3 and 0.4.

layers. To keep the architectures comparable, all the hidden layers have the same number of units and the same dropout rate p. We set the values of $p = 0$ (SGD), 0.1, 0.2, 0.3, 0.4 and the number of epochs $T = 60000$. We observe that the more we increase the number of hidden layers the bigger is the overfitting phenomena.

The authors Srivastava et al. (2014) claim that the choice of p is coupled with the choice of number of hidden units, a high value of dropout rate requires big number of hidden units. To see this effect using our data set, we build 3 single layer perceptrons with 50, 400 and 800 hidden units and set the number of epochs $T = 30000$. The test error of these architectures is given in Figure 2.3-5. We see that the more hidden units we add, the more the SGD curve overfits and the more we need a high dropout rate. This confirms the link between the dropout rate and the number of hidden units.

From our experimentation, to avoid overfitting for a NN, it is better to have a small number of hidden layers with a small number of units (between 50 and 100) and a dropout rate between 0.1 and 0.2. As long as we can't write the explicit expression of the NN dropout objective function, we can't define the best p or the link between p and NN architecture.

Random-forest

In our experimentation, we build 100 maximal trees on which we apply the dropout rates $p =$ 0, 0.1, 0.2, 0.3 , 0.5, 0.66, 0.8 and 0.99. We use the out-of-bag (OOB) prediction values to evaluate the prediction error. In Figure 2.3-6 we see that by dropping out the variables in building the trees, we perform better than the training MSE and the more we increase the dropout rate, the

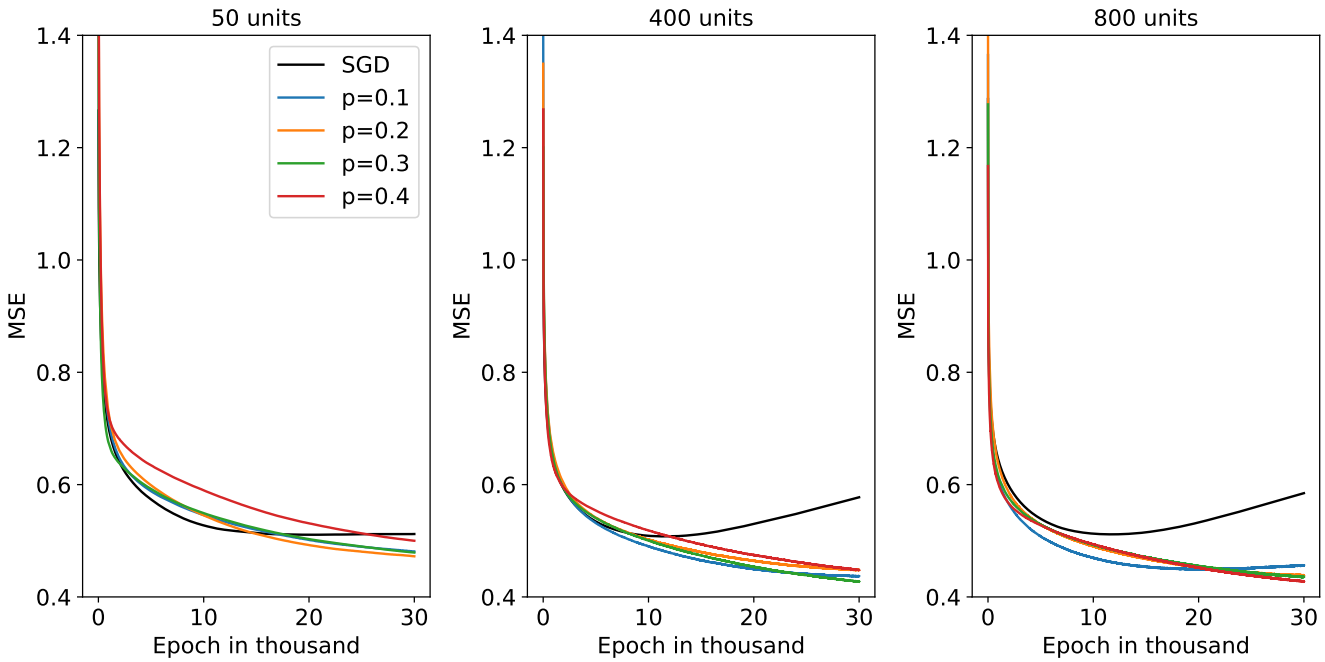

Figure 2.3-5 – Test error of single layer perceptron with a dropout rate $p = 0$ (SGD), 0.1, 0.2, 0.3 and 0.4. Respectively from left to right 50, 400, 800 units in the hidden layer.

smaller is the gap between the training and test MSE. In practice, the default setting in R and Python are $m = d/3$ (see Breiman (2001)) or $m = d$ (see Geurts et al. (2006)) which is equivalent to $p = 2/3$ or $p = 0$. Using our data set, we see in Figure 2.3-6 that a dropout $p = 0$ ($m = d$) seems to be the best choice. We observe that by applying a dropout rate between 0.1 and 0.4, we will not loose much in accuracy.

Prediction procedures comparison

In Table 2.3.2, we give the best result obtained for each procedure. The smallest MSE observed is for single layer perceptron with a dropout rate $p = 0.4$. The random-forest with 100 trees performs as well as the single layer perceptron with 400 units.

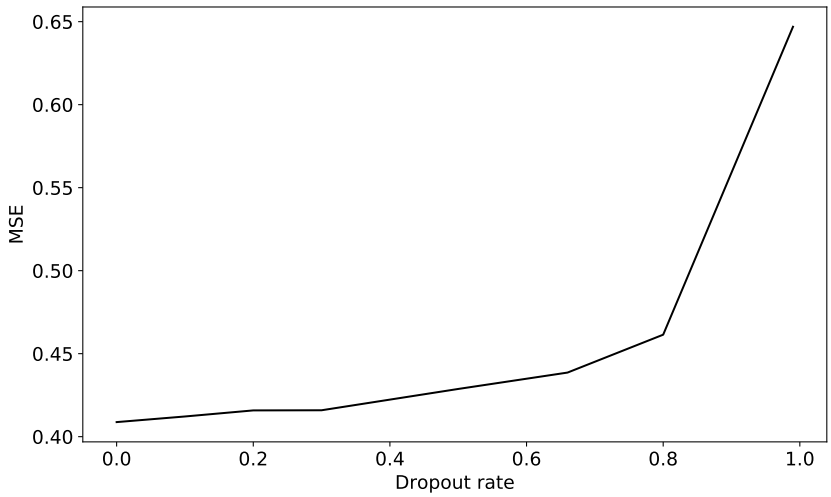

Figure 2.3-6 – Random-forest OOB error thanks to dropout rate p.

Table 2.3.2 – Summary of the best dropout rate by prediction procedure.

Procedure	p	**MSE test**
Linear regression	0.2	0.621
Random-forest 100 trees	0.0	0.406
Single layer perceptron 50 units	0.2	0.444
Single layer perceptron 400 units	0.3	0.405
Single layer perceptron 800 units	0.4	**0.346**
Two layer perceptron 400 units	0.2	0.401
Three layer perceptron 400 units	0.2	0.393

Chapter 3

Functional approach to predict generator oil temperature

This chapter is a reprint of the proceeding *Anomaly detection for aircraft electrical generator using machine learning in a functional data framework* developed by Boulfani et al. (2020) that have been published and presented in the international Global Congress on Electrical Engineering (GC-ElecEng) on September 2020. This work is an application of some methods introduced in Chapter 2 to predict generator oil temperature.

Anomaly detection for aircraft electrical generator in a functional data framework

Fériel Boulfani, Xavier Gendre, Anne Ruiz-Gazen and Martina Salvignol.

Abstract

To reduce the number of aircraft on ground, the electrical design engineers are interested in predicting the oil temperature of the generator during a flight. Changes on the temperature value may indicate an incorrect functioning of the generator. An abnormal behavior can be identified by using machine learning algorithms that predict the generator oil temperature and are trained on flights free from any anomalies. The predictions resulting from the algorithm can then be compared to the observed values, here the sensor data collected from the aircraft during flight. If the observed value is far from the predicted value, a failure warning is raised and a maintenance action shall be performed.

In this paper, we build a digital twin of the electrical generator which predicts the oil generator temperature at a given time thanks to the history of features. We compare several machine learning procedures and the most promising procedure is chosen to predict the generator oil temperature.

The digital twin is tested by using real flight data containing generator failures and it is verified that the algorithm is able to detect an anomaly prior to the failure events (early failure detection).

Keywords

machine-learning; failure detection; generator oil temperature prediction; digital twin; health monitoring;

3.1 Introduction

Several electrical generator failures can lead to a No-Go case, meaning that the aircraft is not allowed to take-off until the failure is fixed. The aircraft has the status of "Aircraft On the Ground" (AOG). The electrical design engineers want to reduce this cost by detecting the abnormal behavior before it turns into AOG and suggest to perform a maintenance action on the generator. Instead of doing a periodic maintenance inspection (preventive maintenance) that immobilizes the aircraft, inspection is performed only when required (predictive maintenance).

The oil circuit cools the generator and its temperature may be used as a measure of proper functioning of the generator. An overheating or a very low oil temperature may indicate a generator anomaly. To monitor the health of the generator (health monitoring), a virtual model that distinguishes anomalies from normal behavior is built. This model can be seen as a Digital Twin (DT) that describes the normal behavior of the electrical generator based on the generator oil temperature. The oil temperature data recorded during the flight represents the physical model that describes the real health of the generator, while the predicted oil temperature represents the virtual model thus the DT.

The idea is to train the DT algorithm using only flights free of anomalies. The prediction of the oil temperature will diverge from the real value when the generator behaves abnormally. Figure 3.1-1 shows how the DT operates to detect an abnormal behavior on simulated flights. The plot of Figure 3.1-1 gives the tracking of the prediction error per flight. The flights are ordered in time, the flight number 50 corresponds to a generator failure. We see that the prediction error starts to diverge 7 flights before the generator failure. An anomaly warning can be raised after the flight number 45 to launch a second DT that investigates the anomaly cause and to select the appropriate maintenance actions. In this paper we will focus on the anomaly detection DT only.

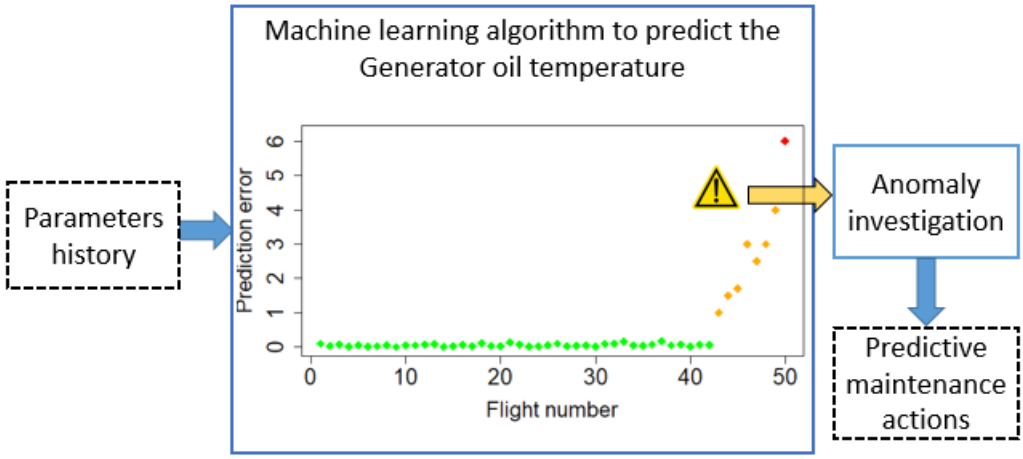

Figure 3.1-1 – Predictive maintenance using DT for a simulated flights with anomaly

The DT concept is widely used in the area of industrial systems maintenance as it saves costs by performing the maintenance only when a warning is raised. A recent state of the art and gap analysis is performed by Aivaliotis et al. (2019) where various applications of the DT concept and its implementation for the maintenance in industrial manufacturing are presented. The authors give the two main reasons why the DT is used which are design validation and product mastering. Our DT model is based on product mastering where the product is the electrical generator.

The reference cited by Aivaliotis et al. (2019) relating to the DT concept for design and predictive maintenance in the aeronautic field is the work of Tuegel (2012) to design and maintain the U.S. Air Force aircraft. The authors used a sub-model for different sub-systems and statistical methods to estimate uncertainty for each sub-model. Moreover in the work of Papachatzakis et al. (2007) and Quintana et al. (2010) the authors calculate the remaining useful life of a system using several methodologies based on the survival analysis. On top of that we add the recent work done by Bachelor et al. (2019) to design and monitor the ice protection system under IT infrastructure constraints.

Our approach differs from the ones mentioned above. To build our DT we decided not to use any physical equations that describe the behaviour of the electrical generator, but to rely only on data measured in service. In fact, the electrical equipment in aeronautics becomes more and more complex, which makes it difficult to build a reliable generator model based on physical knowledge. The aim of this DT is to allow system designers to identify failures by combining their high level knowledge of the electrical system equipment and the data coming from the aircraft.

To build our DT, we choose a functional data framework and predict the oil temperature at time T thanks to the T last records of selected features. This choice is motivated by the high time dependency between the recorded values and adapted to capture the history of the features.

Table 3.2.1 – Features used to predict the generator oil temperature

Parameter	**Description**	**Unit**
y	Generator oil temperature	C°
X^1	Engine speed	Knot (kts)
X^2	Static air temperature	C°
X^3	Total air temperature	C°
X^4	Computed air speed	kts
X^5	Altitude	ft
X^6	Generator load	KVA

The Functional Data Analysis (FDA) encompasses the analysis and theory for functional data Ramsay and Silverman (2005). Such data include data recorded during a time interval. In this paper we use FDA tools to reduce the dimension by expanding our functions onto orthonormal bases and keep the first coefficients of each function. By reducing the dimension we reduce also the model complexity, but decrease the model accuracy. A trade off between complexity and accuracy needs to be done.

The commonly used machine-learning procedures detailed in James et al. (2013) were applied to the coefficients to predict the oil temperature. The procedure that has the smallest prediction error is selected and implemented in the DT.

In this paper we explain the functional data representation in Section 3.2, we recall the prediction procedures and the model selection in Section 3.3. Finally, we test the DT on real flights containing a failure event in Section 3.4. A perspective of this work is given in the concluding Section.

3.2 Functional data

We consider 6 aircraft and we sample $N = 606$ flights of distinct lengths. These flights were checked to be free of anomalies. For each flight, we observe the features defined in Table 3.2.1. The sampling rate of the records is one record per second.

For a given flight $\ell \in \{1, \dots, N\}$, we have n_ℓ observations of $(y, X^1, \dots, X^q)$ where y stands for the oil temperature and $X^1, \dots, X^q$ are the observed values taken by the features. The goal is to predict the oil temperature at a given time thanks to the observations of $X^1, \dots, X^q$ during the T last seconds. For that, we need to split each flight into segments of length T, where $T = \min\{n_1, \dots, n_N\}$. Then, for a flight $\ell \in \{1, \dots, N\}$ we obtain τ_ℓ segments where $\tau_\ell = \lfloor n_\ell / T \rfloor$ and $\lfloor . \rfloor$ stands for the floor part function.

Then for a given flight ℓ, the resulting segmented flight is given by

$$\begin{bmatrix} \dots & X^j(1) & \dots & X^j(T) & \dots \\ \dots & X^j(T+1) & \dots & X^j(2T) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & X^j(\tau_\ell - 1)T + 1) & \dots & X^j(\tau_\ell T) & \dots \end{bmatrix}$$

with τ_ℓ rows and qT columns. The output vector is defined by $y = (y(T), y(2T), \dots, y(\tau_\ell T))$. Figure 3.2-2 shows a visual example of the engine speed feature segmentation.

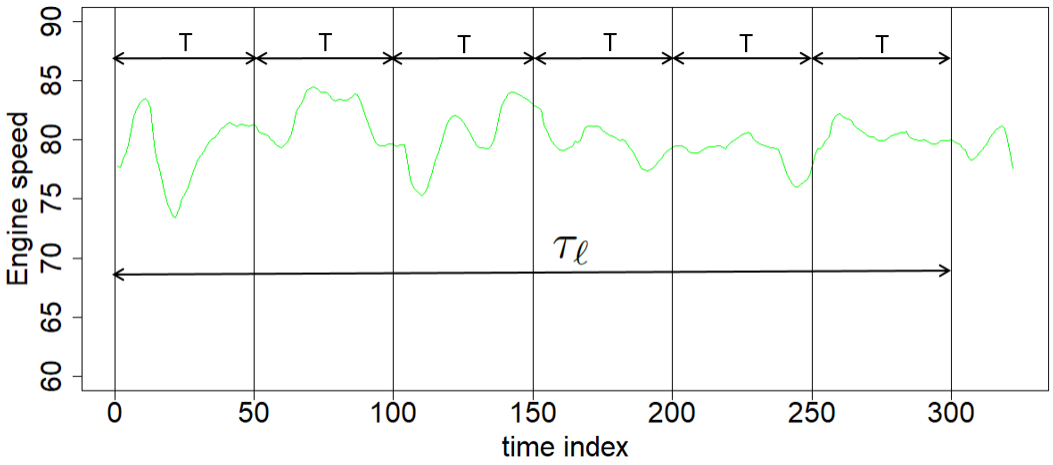

Figure 3.2-2 – Split of engine speed feature into τ_ℓ segments of length T for the flight ℓ

Associated to the discrete observations of variable $X^j, j \in \{1, \dots, q\}$ we consider, for any $j \in \{1, \dots, q\}$ and $k \in \{1, \dots, \tau_\ell\}$, the function :

$$\begin{aligned} x_k^j : [0,1] &\longrightarrow \mathbb{R} \\ t_i &\longrightarrow x_k^j(t_i) = X^j((k-1)T + i) \end{aligned}$$

with $t_i = \dfrac{i}{T}$ where $i \in \{1, \dots, T\}$.

As shown in Table 3.2.1 we are handling features with different units and scales. To normalize the data, we center and scale the functions. For a given function $x^j, j \in \{1, \dots, q\}$, we introduce its average function $\bar{x}^j$ defined by

$$\bar{x}^j(t) = \frac{1}{n} \sum_{i=1}^{n} x_i^j(t), \ t \in \{1, \dots, T\}$$

where n is the row number of the segmented matrix for all flights ($n = \tau_1 + \dots + \tau_N$). The distance

between x^j and its average $\bar{x}^j$ is given by

$$\left\|x_i^j - \bar{x}^j\right\|^2 = \frac{1}{T}\sum_{t=1}^{T}\left(x_i^j(t) - \bar{x}^j(t)\right)^2,\ i \in \{1,\ldots,n\}.$$

Thus, its variance is equal to

$$\mathrm{Var}(x^j) = \frac{1}{n}\sum_{i=1}^{n}\left\|x_i^j - \bar{x}^j\right\|^2.$$

In what follows, we consider all functions centred and scaled as given by

$$\frac{x^j - \bar{x}^j}{\sqrt{\mathrm{Var}(x^j)}},\ j \in \{1,\ldots,q\}.$$

Let $\{\xi_d\}_{d\in\mathbb{N}}$ be an orthonormal basis of the space $L^2 = L^2([0,1], dt)$ of square integrable functions on $[0,1]$ with respect to Lebesgue measure dt. We consider the inner product :

$$\forall f, g \in L^2, \langle f, g\rangle_{L^2} = \int_0^1 f(t)\, g(t)\, dt.$$

The expansion of a function $f \in L^2$ according to the orthonormal basis is :

$$\forall t \in [0,1], f(t) = \sum_{d\in\mathbb{N}} c_d\, \xi_d(t),$$

with $c_d = \langle f, \xi_d\rangle_{L^2}$.

In our application, only discretization of our auxiliary functions are observed. The c_d are estimated on a regular grid of step $1/T$ in $[0,1]$ by

$$\hat{c}_d = \frac{1}{T}\sum_{i=1}^{T} f\left(\frac{i}{T}\right)\, \xi_d\left(\frac{i}{T}\right).$$

There are many ways to represent functional data. In this paper we consider the most used ones, Fourier and Haar wavelet bases Ramsay and Silverman (2005). To reduce the dimension, we plan to handle truncation by selecting the first $D \in \mathbb{N}$ coefficients for the features $X^1, \ldots, X^q$. This means that we have the same number of coefficients D per auxiliary function.

On the matrix C of the coefficients $\hat{c}_d, d \in \{1,\ldots,D\}$, we apply machine learning procedures for regression prediction based on Dq variables instead of Tq with $D \ll T$.

To satisfy orthonormality constraints and to stick to industrial practices, the length T is set to

$256 = 2^8$ seconds in the sequel.

3.3 Prediction procedures

In this section we discuss the statistical procedures that we select to predict the generator oil temperature for regression prediction model defined by

$$y = f(C) + \epsilon,$$

where f is the prediction procedure, ϵ is $n \times 1$ error vector and C is $n \times Dq$ coefficients matrix.

We compare the accuracy of the Neural Network (NN), Ridge Regression (RR) and Random Forest (RF) procedures using the Mean Squared Error (MSE) criterion given by

$$\text{MSE} = \frac{1}{n} \left\| y - \hat{f}(C) \right\|^2,$$

where $\|.\|^2$ is the $\ell 2$-norm defined by : $\|x\|^2 = \sum_{i=1}^{n} x_i^2$.

Hereafter we recall the regression prediction procedures that we use:

- **Ridge Regression** (Friedman et al., 2001, Chap 3): assumes a linear relationship between the input matrix C and the output y with the model $y = CB + \epsilon$. The ridge regression regularizes the MSE by adding a penalty term and the parameters β are estimated by

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|Y - C\beta\|^2 + \lambda \|\beta\|^2 \right\},$$

  where $\lambda > 0$ is the shrinkage parameter.

- **Neural Network** (Friedman et al., 2001, chap 11 ; Bergmeir and Benítez Sánchez, 2012): is a collection of neurons connected together to predict the output y. Each connection between two neurons has a weight. We use a multi-layer perceptron with one hidden layer that contains $\lambda \in \mathbb{N}^*$ number of neurons and the logistic activation function. For the output layer we use the identity activation function. The weight matrix is trained using the backpropagation learning.

- **Random-forest** (Friedman et al., 2001, chap 15 ; Svetnik et al., 2003): is a multitude of decision trees that grow with a random selection of features. The trees are trained at the same time with a bootstrap sample from the data set. The output y is the aggregation of all the outputs of the trees. The number of trees to grow is set at 20 and the number of features to select is denoted by $\lambda \in \mathbb{N}^*$.

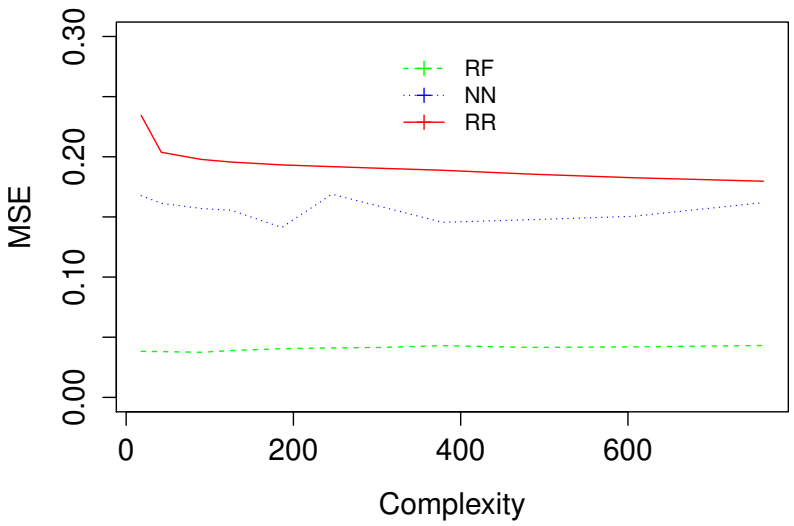

Figure 3.3-3 – Prediction accuracy versus model complexity for the Fourier and Fourier bases

The hyperparameter λ of each procedure needs to be calibrated. For that, we use a common approach known as 10-fold cross-validation (James et al., 2013, Chap 5).

To compare our prediction procedures, we need first to set the number D of selected coefficients for the features. The choice of D is crucial, because the higher D the higher the model complexity. Thus, we need to balance between model performance and model complexity to find an optimal D. Moreover, D is also limited by the features history length T ($D \ll T$).

To set the optimal D, we train the selected procedures with 85% of random flights and keep the 15% to test the accuracy of the procedures. In Figures 3.3-3 and 3.3-4 we plot the link between the model performance MSE and the model complexity which is equal to Dq using Fourier and Haar basis coefficients.

In Figures 3.3-3 and 3.3-4 we differentiate the prediction models RR, NN and RF by respectively solid, dashed and doted lines. This comparison shows that the RF is the most accurate model followed by NN and RR models. For all models and both bases, the MSE increases for a complexity higher than 200. To keep Haar and Fourier basis comparable, we set $D = 31$ and have a complexity $Dq = 186$. By using Fourier and Haar bases expansion we reduce our model from 1 536 to 186 parameters to estimate.

Table 3.3.1 gives the results of the 10-fold cross-validation MSE for $T = 256$, $D = 31$ in order to compare the procedures accuracy. For both bases we obtain similar cross-validation MSE for all procedures. Once again, the RF approach is the best method. In what follows we focus on the NN and RF models and use the Fourier basis only.

To compare NN and RF models performance we plot in Figures 3.3-5 and 3.3-6 the average MSE per flight respectively for the NN and RF model. The flights are separated by aircraft using

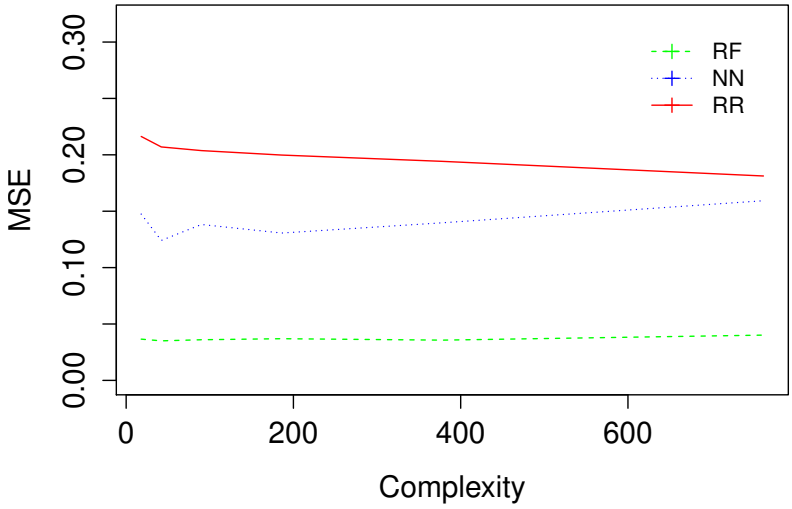

Figure 3.3-4 – Prediction accuracy versus model complexity for the Fourier and Haar bases

Table 3.3.1 – Cross-validation error for RR, NN and RF by basis

MSE	**Fourier basis**	**Haar basis**
RR	0.19	0.20
RF	0.04	0.04
NN	0.10	0.13

vertical lines. The ones used in the training are represented by diamonds and those used in the test by crosses. Horizontal lines (distance mark) are added at the MSE (solid line), 2× MSE (dotted line) and 4× MSE (dashed line) values on both figures.

The MSE per flight for the NN are more centred around the global MSE than the RF. It is noticeable that the RF model suffers form overfitting as the MSE of the test flights are far from the training set. Thus in what follows we keep the NN model and Fourier basis to detect the anomalies.

From Table 3.3.1, we define a reference $\text{MSE}_{\text{ref}} = 0.10$ for NN model using Fourier basis. This value is used in the next section as reference to detect anomalies.

3.4 Anomaly detection using digital twin

To validate that our DT can detect anomalies we need to test it with flights that contain a real anomaly. Unfortunately the anomalies are not easy to identify or to label, and the failures

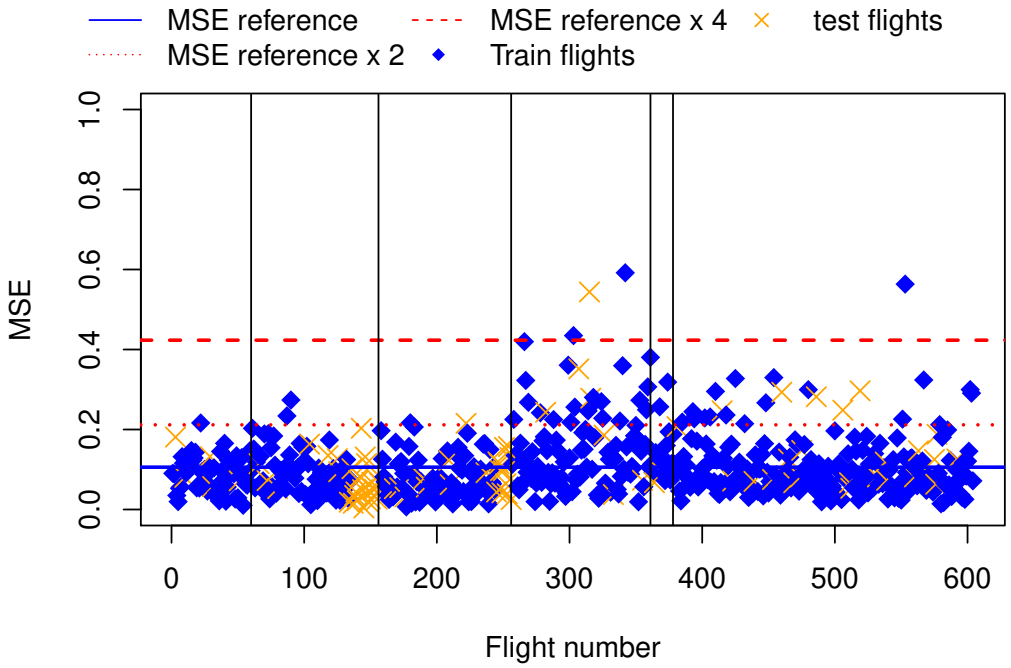

Figure 3.3-5 – Prediction of the generator oil temperature for the flights test using the NN model and Fourier basis

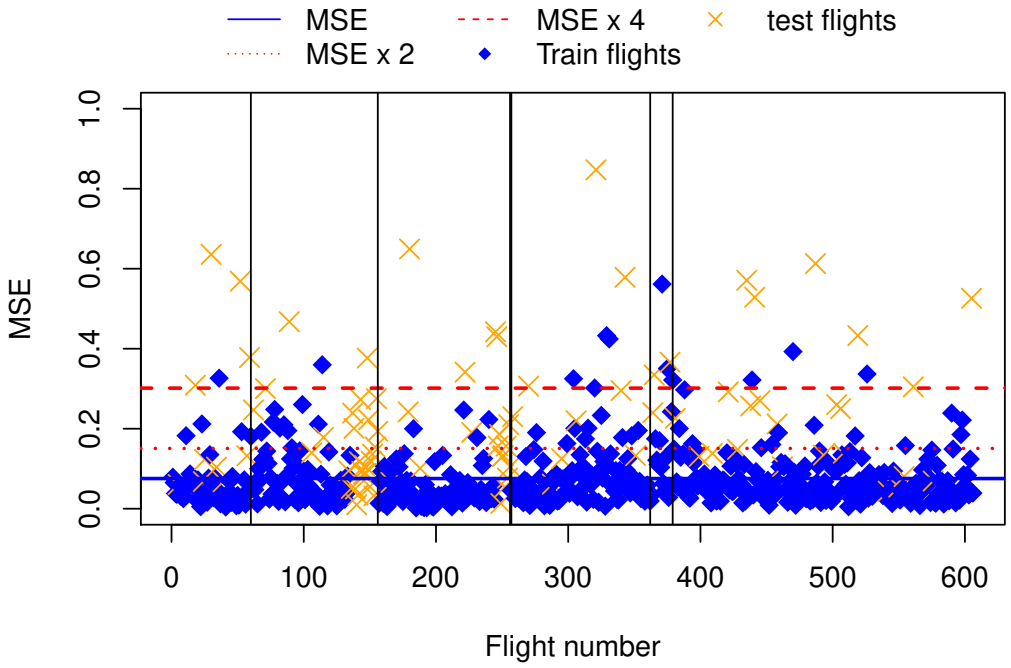

Figure 3.3-6 – Prediction of the generator oil temperature for the flights test using the RF model and Fourier basis

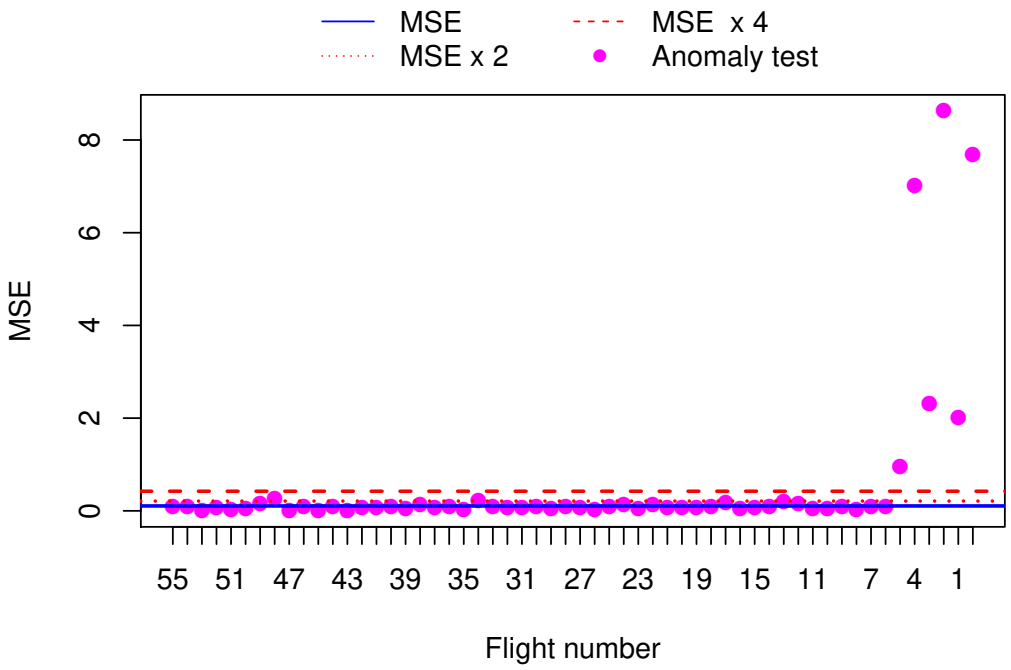

Figure 3.4-7 – MSE by flight to detect anomaly of case 1

occurrence is very rare. We identified two cases of generator failures which represent a loss of one generator in cruise phase in our data set.

For each case, the DT was tested on the flights preceding the generator loss and the results are presented in Figures 3.4-7 and 3.4-8. The distance markers used in Section 3.3 are kept and the flights preceding the generator loss are presented by circles. The flights are numbered using a countdown to the failure, thus the flight number 0 represents the failure.

In case 1 (Figure 3.4-7), the DT is tested on 55 flights preceding the generator loss. The DT starts to detect anomalies 5 flights before it turns into a failure. In Figure 3.4-9 we put all the flights used on this aircraft to benchmark the anomaly test flights and the training flights. It shows that we are predicting well normal behavior and abnormal behavior is identified by an important divergence between prediction and real values.

For the second case (Figure 3.4-8), we have less flights available but the DT is able to detect the failure 9 flights before the failure event.

In these two cases we are facing a huge gap between the MSE of abnormal flights and the distance markers. To identify an anomaly profile, we propose to define a limit defined as a ratio of the MSE of an observed flight over MSE_{ref}. If this limit is exceeded several times in consecutive flights, a warning is raised.

The limit needs to be chosen carefully, as a too high limit may lead to a DT that misses a lot

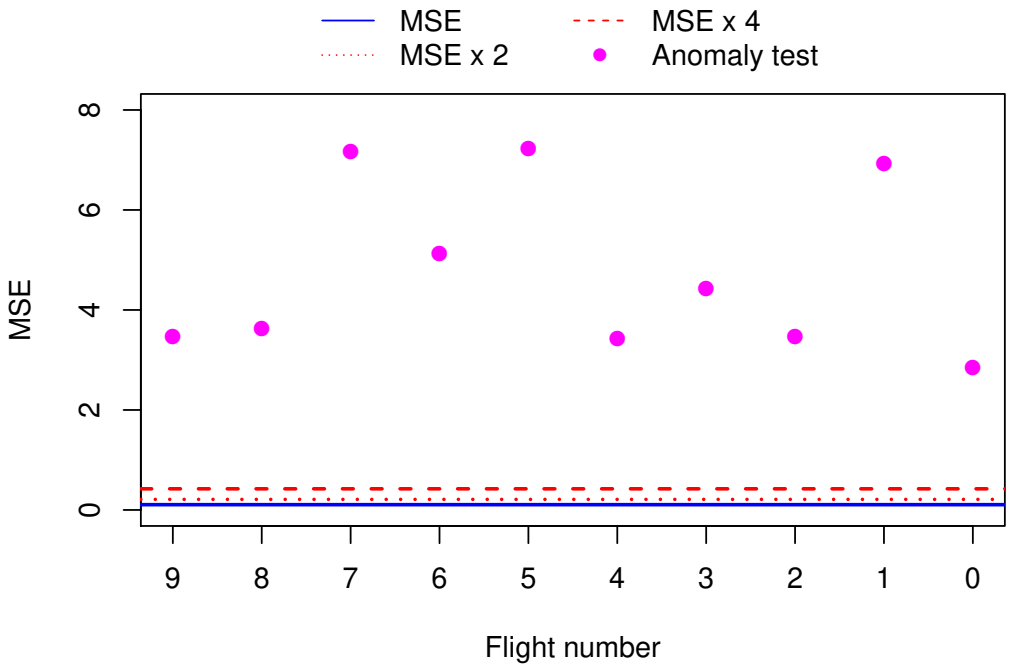

Figure 3.4-8 – MSE by flight to detect anomaly of case 2

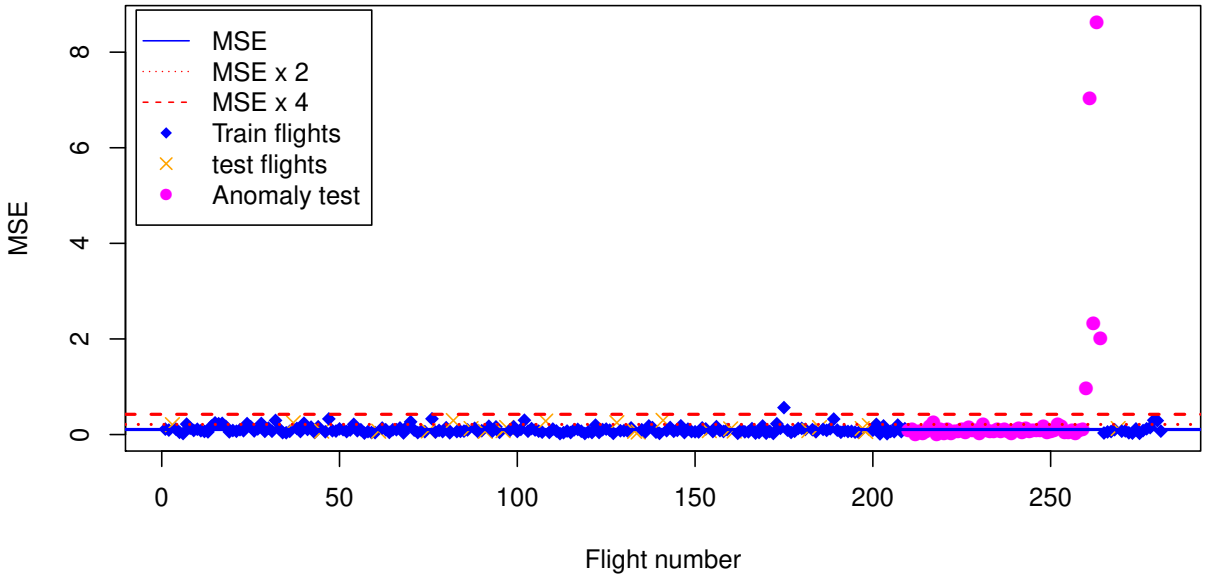

Figure 3.4-9 – MSE by flight to detect anomaly of case 1

of anomalies and a too low limit to a DT that may detect a lot of false anomalies. From the 670 flights with a normal behavior used in this DT the maximum observed MSE by flight is 0.6, thus a limit of 1 is reasonable which represents 10 times MSE_{ref}.

3.5 Conclusion

In this paper we present a way to detect generator anomalies before they turn into failures. We build a digital twin model based on the predicted generator oil temperature value. We use the functional data theory and machine learning procedures to predict the oil temperature thanks to the history of features. The procedures are compared and the best one is selected as a part of the digital twin.

We test the DT built on two cases with anomalies and it shows that the DT is able to detect anomalies several flights before it turns into failure. The amount of data available for this study is not enough to conclude on which different failure types can be detected by this method. However, as AGOs are very costly for airlines and aircraft manufacturers, the here proposed DT has an interest even if it allows to predict a few percentage of failure events.

The main challenges to build this DT was to label the flights correctly to normal and abnormal flights. To improve the anomaly profile, we propose to increase the number of flights with anomalies and to design a limit that will take into account the history of the detected anomalies.

A perspective of this work is to reinforce the learning by injecting the normal flights into the DT and the abnormal flights detected to the anomaly profile to update the limit. The DT could also be adjusted to the real behavior model by adding the generator components remaining lifetime to raise a warning. The warning should raise the probability and the remaining time to get a failure.

Conclusion

In this part, we explain how to use linear regression, neural network and random-forest procedures in a regression context when the explanatory variables are functions.

In Chapter 2, we study the dropout regularization as a way to prevent overfitting phenomena and we analyse the effect of dropout regularization on the accuracy of the procedures. We write an explicit expression of dropout regularization term for linear model and single layer perceptron. We show that the dropout on single layer perceptron acts as a Tikhonov regularizer with a parameter $p(1-p)^{-1}$ and is related to the empirical second moment of the outputs of hidden neurons. Thanks to some experimentation, we found that the higher is the number of hidden units the higher should be the dropout rate.

One perspective of the dropout regularization which needs further development is to provide an efficient way to predict the oil temperature for a range of data that is unobserved (extreme conditions) and compare the manufacturer simulated values with the predicted values.

Chapter 3 uses the functional data theory to predict the abnormal behavior of a generator. The objective is to reduce the number of aircraft on ground, by predicting the oil temperature of the generator during a flight. This work was presented at the international Global Congress on Electrical Engineering in September 2020. The approach consists in training some machine learning procedures with flights free from any anomalies and in detecting abnormal flights as the ones with large prediction errors. Several machine learning methods are compared.

We integrate this approach in a digital twin that detects a generator failures and we test it on flights that contain failures. We illustrate on some real data set that this digital twin is able to detect some anomaly behavior prior to the failure event.

Part III

Abnormal behavior detection of aeronautical electrical generator

Sommaire

Introduction

Anomalies or outliers detection process is used to identify unusual items or events in data sets. In the aeronautic field, with high reliability standards, the Invariant Component Selection (ICS) method is relevant since it can be adjusted to data sets with a small percentage of outliers (less than 5%).

In multivariate anomaly detection, Archimbaud et al. (2018) consider the problem of outlier detection in large dimension and propose to use ICS. In our context, the electrical generator behavior is observed on a very fine time scale during several flights and a dozen of technical parameters such as airspeed, altitude, pitch, roll, etc. are recorded throughout flights by an aircraft. This motivates the choice of adapting ICS for anomaly detection in the functional data framework. This ICS generalization is the object of the present part.

In Chapter 4, we use several data sets which contain a several functional variables to explain the generalization of ICS in the multivariate functional anomaly detection framework. The selected data sets are from aeronautic, in a predictive maintenance context, and from manufacturing, in a quality control context. Using these data sets, we discuss the adaptation of the ICS method from the multivariate to the multivariate functional case. Two approaches of ICS are proposed in the functional framework: point-wise and global. The outliers identification and interpretation are discussed for both approaches and both data sets.

In this part, we use the notations from Archimbaud (2018b). A p-multivariate dataset containing n observations is denoted by $\mathbf{X}_n$ and the p-multivariate vector associated to observation i by $\mathbf{x}_i$.

Chapter 4

ICS for multivariate functional anomaly detection

This chapter is a reprint of the paper *ICS for multivariate functional anomaly detection with applications to predictive maintenance and quality control* that has been submitted to the Journal "Econometrics & Statistics".

ICS for multivariate functional anomaly detection with applications to predictive maintenance and quality control

Aurore Archimbaud, Fériel Boulfani, Xavier Gendre,
Klaus Nordhausen, Anne Ruiz-Gazen and Joni Virta.

Abstract

Multivariate functional anomaly detection has received a large amount of attention recently. Accounting both the time dimension and the correlations between variables is challenging due to the existence of different types of outliers and the dimension of the data. Most of the existing methods focus on a small number of variables. In the context of predictive maintenance and quality control however, data sets often contain a large number of functional variables. Moreover, in fields that have high reliability standards, detecting a small number of potential multivariate functional outliers with as few false positives as possible is crucial. In such a context, the adaptation of the Invariant Component Selection (ICS) method from the multivariate to the multivariate functional case is of particular interest. Two extensions of ICS are proposed: point-wise and global. For both

methods, the choice of the relevant components together with outlier identification and interpretation are discussed. A comparison is made on a predictive maintenance example from the avionics field and a quality control example from the microelectronics field. It appears that in such a context, point-wise and global ICS with a small number of selected components are complementary and can be recommended.

4.1 Introduction

Currently, multivariate functional data are frequently encountered in many different fields including meteorology (Suhaila et al., 2011), medicine (Erbas et al., 2007) and quality control (Millán-Roures et al., 2018). More precisely, the observations that we consider are functions of a univariate input variable (usually time) with multivariate output values. In what follows, the number of dimensions of the vector of output curves is denoted by p and the number of observed sets of these p curves is denoted by n. Daily measurements of temperature, log precipitation and wind speed at some weather stations provide one example of such data. Other examples come from high reliability standards field, such as automotive, avionics or aerospace where many parameters are measured over a certain period of time. In avionics, dozens of technical parameters, such as the airspeed, altitude, and so on, are recorded throughout flights by aircraft. In microelectronics semiconductor fabrication, there is a large collection of process-control measurements, which are also recorded from various sensors during the processing of silicon wafers.

In the usual non-functional multivariate framework, it is well-known that anomalies might not be outlying in any of the original variables but could exhibit a different correlation pattern compared to the main bulk of the data. While univariate outliers can be identified quite easily, specifically by visually looking for extreme values, multivariate outliers are more difficult to detect, especially in large dimension. In the univariate functional context, outliers in the sense of their magnitude are usually distinguished from outliers in the sense of their shape. While magnitude outliers can be identified visually by looking at the extreme values of the curves, it can be difficult to detect observations that do not have especially high nor low values but instead have different patterns. When looking at multivariate functional data with a large number of dimensions p, both difficulties (multivariate and functional) are combined, and there are many possible types of outliers (see Hubert et al. (2015) for a taxonomy). Outliers are usually first divided into isolated or persistent outliers depending on whether their abnormal behaviour last for a short or long time. In the present contribution, we focus on high reliability applications; we assume that an extreme behaviour over a very short time is detected beforehand, and we focus on outliers that are quite persistent.

Detecting outliers in a multivariate functional framework is an issue that has received a large amount of attention very recently (see Rousseeuw et al. (2018), Staerman et al. (2019), Dai et al.

(2020), Lejeune et al. (2020) and the references therein). Given that shape outliers are more difficult to identify than magnitude outliers, several recent papers tackle the problem of detecting shape anomalies in either a univariate functional framework (Nagy et al. (2017) and Harris et al. (2020)) or a multivariate functional framework (Dai et al. (2020) and Lejeune et al. (2020)).

Many papers in the univariate and multivariate functional data analysis literature consider the problem of outlier detection through a functional depth or pseudo-depth approach (e.g. Hubert et al. (2015), Kuhnt and Rehage (2016), Dai et al. (2020) and references therein). Each depth notion leads to a centrality index for the observed curves that allows for the identification of non central curves as outliers. In a multivariate context with p large ($p \geq 5$), the depth-based methods are computationally costly. There exist, however, other approaches such as the shape-based features extraction method by Lejeune et al. (2020) and the isolation forest method by Staerman et al. (2019). Nevertheless, none of the proposed methods incorporate a dimension reduction step with regard to the dimension p. The lack of a dimension reduction step most likely explains, besides the computational burden, why most examples discussed in the literature on multivariate functional outlier detection do not go beyond $p = 3$, and they instead focus on the bivariate case (Kuhnt and Rehage (2016), Rousseeuw et al. (2018), Dai and Genton (2019), Dai et al. (2020), Staerman et al. (2019) and Lejeune et al. (2020)).

In a non-functional framework, Archimbaud et al. (2018) consider the problem of outlier detection in large dimension (but still $n > p$) and show that the Mahalanobis distance, which is a particular depth measure, works poorly. They propose to use instead the Invariant Coordinate Selection (ICS) method (see also Archimbaud et al. (2018c) and the R packages ICSOutlier by Archimbaud et al. (2018b) and ICSShiny by Archimbaud et al. (2018a) for the implementation of the method in R). ICS is based on the joint diagonalization of two scatter matrices. The theoretical properties of the method are studied in Tyler et al. (2009) for a mixture of elliptical distributions on the one hand, and for Independent Component Analysis on the other hand. The method is similar to PCA in the sense that it allows a dimension reduction by calculating and selecting a small number of coordinates or components. However, instead of relying on the eigendecomposition of one scatter matrix, it relies on the eigendecomposition of one scatter matrix relative to a second matrix. When observations are structured in groups as is the case in the presence of outliers (small groups), ICS is capable of recovering the Fisher discriminant subspace without knowing the group membership (see Theorem 3 in Tyler et al. (2009)). With regard to outlier detection, this capability means that the dimension reduction through ICS is more likely to retain the outlyingness structure compared to PCA which is not intended to recover the Fisher discriminant subspace. Moreover, ICS is affine invariant while PCA is only orthogonally invariant. Archimbaud et al. (2018) consider different pairs of scatter matrices where one scatter is more robust than the other. In high reliability standards areas, such as automotive, avionics or aerospace where only a small proportion observations can be abnormal, the authors recommend the use of the regular covariance matrix and the so called matrix of fourth moment as the scatter pair. They also exhibit the

advantage of ICS compared with the Mahalanobis distance and with robust PCA.

Recently, Li et al. (2019) and Virta et al. (2020) proposed to generalize ICS to functional data in the context of Independent Component Analysis. While Li et al. (2019) focus on the univariate case, Virta et al. (2020) consider multivariate functional data. To the best of our knowledge, there exists no extension of ICS for multivariate functional outlier detection. In the present paper we propose two functional ICS extensions.

The first method called "point-wise ICS" is comparable with the point-wise approach used in Dai and Genton (2018) and Rousseeuw et al. (2018) but it replaces, at each time step, the depth procedure by ICS. As detailed in Archimbaud et al. (2018), ICS consists of calculating invariant coordinates, selecting the relevant components and calculating, for each observation, an outlyingness score using only the selected components. To select the relevant components, we propose to use the asymptotic test from Nordhausen et al. (2017) at each time point. Following Rousseeuw et al. (2018), the scores obtained are summarized by calculating certain average and dispersion measures. Using the Functional Outlier Map (FOM) from Rousseeuw et al. (2018), the amplitude and shape outliers can be identified. The second ICS adaptation to functional data is called "global ICS" and consists of expanding the data on basis functions, such as Fourier or B-splines, and selecting a number D of these basis functions. ICS is then applied to the vectors of $p \times D$ coordinates. The procedure is very similar to the procedure described in Virta et al. (2020) in the case of Independent Component Analysis when using an orthonormal basis but it must be completed by an outlier identification step. For global ICS, we follow the recommendations from Archimbaud et al. (2018) and use the scree plot as a simple tool to choose the relevant components and identify outliers using a Monte Carlo cutoff. To illustrate and compare "point-wise" and "global" ICS, we propose to examine in detail an example of daily weather measurements in small dimension. This example illustrates that point-wise and global ICS can identify amplitude and shape outliers with a multivariate perspective. Selecting a small number of components allows us to identify the most extreme multivariate amplitude and shape outliers. In the context of quality control, where univariate outliers have already been detected upstream, a small false positive rate is crucial. Thus, point-wise and global ICS with a small number of selected components are interesting approaches as illustrated on two real data sets.

This paper is organized as follows. Section 4.2 is divided into four subsections. In Subsection 4.2.1, we recall ordinary ICS for multivariate outlier detection. In Subsection 4.2.2, we consider multivariate functional data and detail a preprocess that is used in deriving equispaced data and reducing the dimension by projecting the data onto functional bases. Then, we propose to generalize ICS to multivariate functional outlier detection in two different ways. Point-wise functional ICS consists of applying the ordinary ICS at each time point and is detailed in Subsection 4.2.3. Global ICS consists of implementing ICS only once on the coefficients of a functional basis expansion of each of the variables and is detailed in Subsection 4.2.4. Details concerning

the criterion for the dimension choice and the cutoff for outlier detection are also provided in Subsections 4.2.3 and 4.2.4 together with an illustration using the weather data set. Section 4.3 illustrates the two functional ICS approaches on two real data sets from the predictive maintenance and quality control fields. Finally Section 4.4 concludes and gives perspectives.

4.2 ICS for multivariate functional data

4.2.1 Ordinary ICS

For a p-variate data set $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ where $'$ denotes the transpose operator, a scatter matrix is a $p \times p$ symmetric positive definite and affine equivariant matrix. We recall that an affine equivariant matrix $\mathbf{V}(\mathbf{X}_n)$ is such that

$$\mathbf{V}(\mathbf{X}_n\mathbf{A} + \mathbf{1}_n\mathbf{b}') = \mathbf{A}'\mathbf{V}(\mathbf{X}_n)\mathbf{A},$$

where $\mathbf{A}$ is a full rank $p \times p$ matrix, $\mathbf{b}$ a p-vector and $\mathbf{1}_n$ an n-vector full of ones. One scatter pair example is the regular covariance matrix

$$\mathrm{COV}(\mathbf{X}_n) = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

where $\bar{\mathbf{x}}$ denotes the empirical mean, and the so called scatter matrix of fourth moments

$$\mathrm{COV}_4(\mathbf{X}_n) = \frac{1}{(p+2)n}\sum_{i=1}^{n} r_i^2(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})',$$

where $r_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathrm{COV}(\mathbf{X}_n)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})$ is the classical squared Mahalanobis distance (see Nordhausen and Tyler (2015) and the many references therein for other scatter matrices examples). As illustrated on simulations by Archimbaud et al. (2018) in the context of a small proportion of outliers, this particular scatter pair is not only simple and fast to compute but also efficient in detecting outliers when compared to other pairs that involve robust scatter estimators. The context in which there is a small proportion of outliers (less than 2%) is encountered in fields where the data quality standards are high, such as in avionics or aerospace fields, and it is the context that we are interested in.

ICS consists in the joint diagonalization of a scatter pair $\mathbf{V}_1(\mathbf{X}_n)$ and $\mathbf{V}_2(\mathbf{X}_n)$. It leads to a $p \times p$ matrix $\mathbf{B}(\mathbf{X}_n)$ and a diagonal matrix $\mathbf{D}(\mathbf{X}_n)$ such that:

$$\mathbf{B}(\mathbf{X}_n)\mathbf{V}_1(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{I}_p \quad \text{and} \quad \mathbf{B}(\mathbf{X}_n)\mathbf{V}_2(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{D}(\mathbf{X}_n)$$

where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. $\mathbf{D}(\mathbf{X}_n)$ contains the eigenvalues of $\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)$ in decreasing order, while the rows of the matrix $\mathbf{B}(\mathbf{X}_n) = (\mathbf{b}_1, \ldots, \mathbf{b}_p)'$ contain the corresponding eigenvectors so that:

$$\mathbf{V}_1(\mathbf{X}_n)^{-1}\mathbf{V}_2(\mathbf{X}_n)\mathbf{B}(\mathbf{X}_n)' = \mathbf{B}(\mathbf{X}_n)'\mathbf{D}(\mathbf{X}_n).$$

Using any affine equivariant location estimator $\mathbf{m}(\mathbf{X}_n)$, the corresponding scores

$$\mathbf{Z}_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)' = (\mathbf{X}_n - \mathbf{1}_n\mathbf{m}(\mathbf{X}_n)')\mathbf{B}(\mathbf{X}_n)'$$

where $\mathbf{1}_n$ denotes the n-vector of ones are the affine invariant coordinates or components.

As proved in Archimbaud et al. (2018), the Euclidian norm $\sqrt{\mathbf{z}_i'\mathbf{z}_i}$ of observation $i = 1, \ldots, n$ is equal to the Mahalanobis distance of observation i from the location $\mathbf{m}(\mathbf{X}_n)$ in the sense of $\mathbf{V}_1(\mathbf{X}_n)$. The Mahalanobis distance does not offer the possibility of dimension reduction. However, this property can be useful if the outliers belong to a space of reduced dimension and if we attempt to avoid false positives, as in high reliability fields. In contrast, ICS offers the possibility of selecting the components that are helpful in detecting real anomalies, and consequently, it avoids false positives due to noisy dimensions. In the case of a small proportion of outliers, the theoretical properties of ICS (see Archimbaud et al. (2018) for details) lead us to focus on only the invariant components associated with the largest eigenvalues. Archimbaud et al. (2018) propose different automatic selection procedures based on hypothesis testing, but they acknowledge the fact that these procedures tend to select too many components, and they propose the scree plot as an alternative. Once having selected k invariant components, the last step in the procedure is the identification of the outlying observations. For each observation $i = 1, \ldots, n$, we calculate its squared "ICS distance", which corresponds to its squared Euclidian norm in the invariant coordinate system while accounting for the k first coordinates:

$$(\text{ICS distance})_{i,k}^2 = \sum_{j=1}^{k} \left(z_i^j\right)^2 \tag{4.2.1}$$

where z_i^j denotes the jth coordinate of the score $\mathbf{z}_i$. In Archimbaud et al. (2018), an observation is flagged as an outlier when its ICS distance using k components is larger than a cutoff based on Monte Carlo simulations from the standard Gaussian distribution. Being given a data dimension, a scatter pair and a number k of selected components, many Gaussian samples are generated and the ICS distances are computed. A cutoff is derived for a fixed level γ as the $1 - \gamma$ percentile of these distances.

4.2.2 Preprocessing multivariate functional data

Let assume that our data are vectors of square integrable functions on $[0,1]$ with respect to Lebesgue measure dt. We denote the Hilbert space of such functions by $L^2 = L^2([0,1], dt)$. Note that there is no loss of generality to restrict ourselves to functions on the time interval $[0,1]$. Hereafter, we consider the usual inner product,

$$\forall f, g \in L^2, \langle f, g \rangle_{L^2} = \int_0^1 f(t)\, g(t)\, dt.$$

Let us assume that we have a data set with p variables (or features) and each one has dimension $n \times T$, where n is the number of observed curves or functions and T is the number of observed time points. In practice, the number of time points can differ across both observations and features, and alignment or warping is necessary. For point-wise ICS, we need the curves to be aligned for both observations and features, while for global ICS, the alignment is necessary only across the observations, but different features can have different alignments. In both data analyses of Section 4.3, however, the features are already aligned for each observation and as a result, the curves are only required to be aligned across observations. For the sake of simplicity, we use a linear interpolation. Other methods such as the method proposed by Tucker et al. (2013) and implemented in Tucker (2020) are also possible. In practice, we fix a value for the number of time points T, for example at the median value of the different curve time lengths. Then, we linearly interpolate each curve at T equally spaced points. For a curve with a length greater than T we reduce the number of points by averaging points and for the curves with a length smaller than T, we add points using linear interpolation.

For point-wise ICS, no other preprocessing is necessary. However, for global ICS, a preliminary dimension reduction is necessary. We consider expansion of the functions associated with the aligned data on a given basis. For an orthonormal basis $\{\xi_d\}_{d \in \mathbb{N}}$, the expansion of a function $f \in L^2$ is given by

$$\forall\, t \in [0,1],\ f(t) = \sum_{d \in \mathbb{N}} c_d\, \xi_d(t), \text{ with } c_d = \langle f, \xi_d \rangle_{L^2}.$$

In our application, only discretizations of our functions are observed which leads to an estimation of c_d on a regular grid of step $1/T$ in $[0,1]$ which is given by

$$\hat{c}_d = \frac{1}{T} \sum_{t=1}^{T} f\left(\frac{t}{T}\right)\, \xi_d\left(\frac{t}{T}\right).$$

Among the many possible bases (Ramsay and Silverman, 2005), we consider the Fourier and the B-splines bases which are very commonly used. Note that, contrarily to the Fourier basis,

the B-splines basis is not orthonormal and the coefficients estimation formula has to be adapted in order to take into account the Gram matrix. In what follows, we use the usual least squares approach for the estimation and do not take into account the Gram matrix when applying global ICS (see Subsection 5.1 in Virta et al. (2020) for details on the implementation of ICS taking into account the Gram matrix). To reduce the dimension, a truncation is made, and the first $D \in \mathbb{N}$ coefficients are selected for each variable. The resulting truncation gives a data set with dimension $n \times pD$. We choose the same number of coefficients for each of the variables for reasons of simplicity but it is quite possible to vary this number. If we choose a number D_j of coefficients for each of the variables $j = 1, \ldots, p$, we get a data set of dimension $n \times \sum_{j=1}^{k} D_j$.

As detailed in Barreyre et al. (2019) in the context of satellite data, the choice of the basis to represent functional data in a reduced dimension could have an impact on the whole outlier detection procedure. In particular, the authors exhibit an artificial example in which isolated outliers are more likely to be identified when using a data-dependent basis such as PCA rather than a Fourier basis. However, as mentioned in the introduction, we do not focus on isolated outliers, and instead, we consider Fourier and B-splines bases, which often give similar results in our experience. As mentioned earlier, it would also have been possible to take into account the Gram matrix for the B-splines basis which is not orthonormal.

In what follows, we detail two generalizations of ICS to multivariate functional data. The outlier detection methodologies are fully described and illustrated on a small Spanish weather dataset from the R package *fda.usc*. The data set consists of $p = 3$ variables that represent the daily average ($T = 365$) of temperature, wind speed and log precipitation records from 1980 to 2009 from $n = 73$ weather stations in Spain. As in Dai and Genton (2018) the curves are smoothed using a B-spline basis truncated at $D = 11$. No expert opinion on outlying weather stations is available but the example is small enough in terms of the number of observations and variables to be studied in detail by examining the curves. To help the reader to understand the data, the curves are plotted in the appendix with different colours (see Figures 4.6-23 to 4.6-25). For the interpretation of the colours, the reader is referred to the web version of this paper. More details can also be found in Dai and Genton (2019) and Dai et al. (2020). Each figure focuses on one of the 3 variables. In Figure 4.6-23, we coloured in red and identified by a number the curves that look different from the vast majority of curves in terms of their temperature behaviour. These are curves 56 and 45, which have very low temperatures, and the seven curves 34, 35, 36, 55, 57, 58 and 60, which have much flatter temperature curves than the other weather stations (see the left panel). These nine curves are also coloured in red on the wind speed (resp., log precipitation) plot in the middle (resp., right) panel of the same Figure. In Figure 4.6-24, we have kept the red curves and coloured in blue the curves that were not already coloured in red and that look different from the others in terms of the wind speed. Curves 20 and 59 take large values with several large bumps. Curves 51 and 72 take small wind speed values. Finally, in Figure 4.6-25, we have kept the

red and blue curves and coloured in green the curves that were not already coloured in red or blue and that look different from the others in terms of the log precipitation. These are curves 33, 39, 44 and 66, which have large values of log precipitation with a small dispersion. All together, this process gives 17 curves out of 73 (23%) that can be suspected as outlying. In the present paper, the objective is to detect only a small percentage (approximately 2%) of the observations as outliers, which corresponds to at most two or three observations in this small data set. Clusters of outliers, such as some of the red and green curves in Figure 4.6-24, are not of interest in our context, which means that we are rather looking for observations that differ as much as possible from other observations (while accounting for the interactions between the variables) and that are unique in their outlying behaviour. For both generalizations of ICS to functional data, the different steps of the outlying detection procedure are now detailed, namely, the invariant components calculation, the dimension reduction, the outlier identification and the outlier interpretation. Differences and complementarities of point-wise and global ICS are also discussed.

4.2.3 Point-wise functional ICS

Once the curves are aligned, it is possible to implement an ordinary ICS at each time point $t = 1, \ldots, T$ and calculate p invariant components for each of the T ICS analyses, which means that we have pT components. We then select $k(t)$ components at every time t. For each observation i, we calculate an $(\text{ICS distance})^2_{i,k(t)}(t)$ using expression (4.2.1) at every time t. Note that if there is no dimension reduction ($k(t) = p$), the method is equivalent to the calculus of a Mahalanobis distance at each time point.

The main advantage of ICS distances compared to Mahalanobis distances is that they can be based on a subset of components that form a more informative subspace. Basically all of those components that appear to be the most non-Gaussian should be selected. ICS can be considered in this case to be non-Gaussian component analysis (NGCA) (Nordhausen et al., 2017), where we assume that all of the outliers lie in a subspace that obviously is non-Gaussian and that this subspace is independent from the uncontaminated Gaussian subspace. Tyler et al. (2009); Nordhausen et al. (2017); Radojicic and Nordhausen (2020) then show that the eigenvalues $d_1, \ldots, d_p$ contained in $\mathbf{D}(\mathbf{X}_n)$ are the key to identifying of the two subspaces because these can be seen as generalized measures of kurtosis. For the Gaussian subspace all eigenvalues must thus be equal and the exact values depend on the scatter matrices used in ICS. In the case of COV and COV_4, it can be shown that the "Gaussian" eigenvalues are equal to 1, and in our setting of only a few outliers, the "Non-Gaussian" eigenvalues are all larger than one. Thus, we have $d_1 \geq \ldots \geq d_k > 1 = \ldots = 1$, where the problem is now that k is unknown. One could, for example, use a scree plot or marginal tests as discussed in Archimbaud et al. (2018) or use successive applications of hypothesis tests of the form $H_{0q}: \; k = q$ and test H_{00}, H_{01}, H_{02}, ... to find the value $\hat{k} = q$ where H_{0q} is the first test not rejected at a given significance level. The test statistic and its limiting distribution under

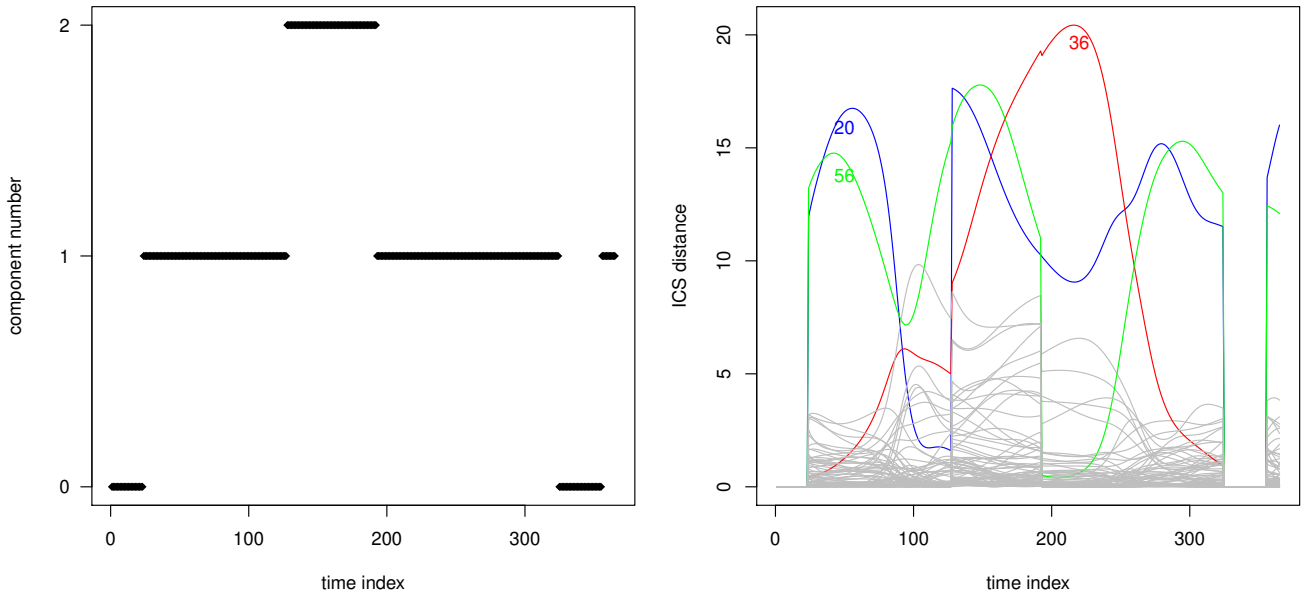

Figure 4.2-1 – **Weather data set** - Point-wise ICS, Left: Number of components selected. Right: square ICS distance.

the null hypothesis for a given q are presented in Nordhausen et al. (2017) where it is also shown that these are fast to compute and lead to a consistent estimate of k for an appropriate sequence of significance levels. In what follows we take 1% as the initial significance level for testing H_{00}, and then, we apply a type of Bonferroni adjustment by dividing the level by 2 for H_{01}, by 3 for H_{02}, and so on (see Archimbaud et al. (2018) for more details). For point-wise functional ICS, this procedure should be performed separately for every time point, which means that the number of selected components varies over time.

At each time point, a subset of components is chosen, and ICS assigns a measure of outlyingness (ICS squared distance) to every observation. It is thus possible to plot the number of selected components on the one hand and the outlyingness measures on the other hand as functions of time. These plots are given in Figure 4.2-1 for the weather data set. When no component is selected, no outliers are detected and the ICS distances are thus equal to 0 for these time points. As mentioned previously, the number of selected components varies over time. However, on Figure 4.2-1 (left panel), the plot is quite structured with only one selected component except for in the middle of the year, where two components are selected. Note that selecting more components usually leads to more outliers. Looking at the ICS distance curves (the right panel of Figure 4.2-1), it is possible to identify the curves 20, 36 and 56 as outlying and detect at which periods of time they differ from the other curves. Observations 20 and 56 differ from the other curves at similar periods of time (at approximately the 50th, 150th and 300th days of the year), but their ICS distances are not the same. Curve 36 differs from the other curves essentially at approximately

the 200th day. In the presence of many curves, it can be tedious to identify outliers by looking at such a plot. Moreover, it is costly and not recommended to flag the outliers by calculating a cutoff based on Monte Carlo simulations at each time point. We thus propose to summarize the information and to flag the outliers by using a functional outlier map (FOM), as defined by Rousseeuw et al. (2018). For each observation, we calculate and plot a weighted average (fICS) and a measure of variability (vICS) for the ICS squared distances:

$$\text{fICS}_i = \sum_{t=1}^{T} W(t) \left(\text{ICS distance}\right)^2_{i,k(t)}(t) \tag{4.2.2}$$

$$\text{vICS}_i = \frac{\text{stdev}_i}{1 + \text{fICS}_i} \tag{4.2.3}$$

where $W(.)$ is a weight function such that $\sum_{t=1}^{T} W(t) = 1$ and stdev_i denotes the standard deviation of the $\left(\text{ICS distance}\right)^2_{i,k(t)}(t)$ values over time. As mentioned in Rousseeuw et al. (2018), vICS is a relative measure that is preferable to the usual standard deviation. We could also use a weighted standard deviation as suggested in Rousseeuw et al. (2018). In the absence of additional information, we limit ourselves to uniform weights $W(t) = 1/T$. Note, however, that the number of components involved in the calculation of the ICS distances could vary from one time to another, and the larger the number of components selected is, the larger the ICS distance. This relationship means that time points with a large number of selected components can have a larger impact in the fICS calculation than time points with a small number of selected components. This relationship is especially true for time points at which no components are selected because the ICS distances are zero and such time points do not contribute to the fICS value. In some sense, the fICS is a data-driven weighted average of the ICS squared distances with no need for auxiliary information. If one wants to give truly uniform weights over time, it is also possible to divide the squared ICS distances at each time by the number of selected dimensions as illustrated in Section 4.3. This standardization is equivalent to take $W(t) = 1/(Tk(t))$ in (4.2.2). The FOM is a scatterplot of fICS and vICS. Large values of fICS correspond to curves that are outlying during a long period or for the entire period of time. Note that such curves are not necessarily shifted curves since a high ICS distance can correspond to a multivariate outlier (an observation outlying in the correlation structure but not necessarily extreme). Large vICS values correspond to curves whose behaviour differs from the other curves during some subperiods of time. Figure 4.2-2 gives the FOM for the weather data set. Note that the cutoff curve (red dashed curve on the Figure) is calculated as in Rousseeuw et al. (2018) using the combined functional outlyingness (CFO) with a quantile order of 0.95. Its calculation is adapted from Rousseeuw et al. (2018). For each observation $i = 1, \ldots, n$, we define

$$\text{LfICS}_i = \log\left[0.1 + \sqrt{\left(\frac{\text{fICS}_i}{\text{med(fICS)}}\right)^2 + \left(\frac{\text{vICS}_i}{\text{med(vICS)}}\right)^2}\right]$$

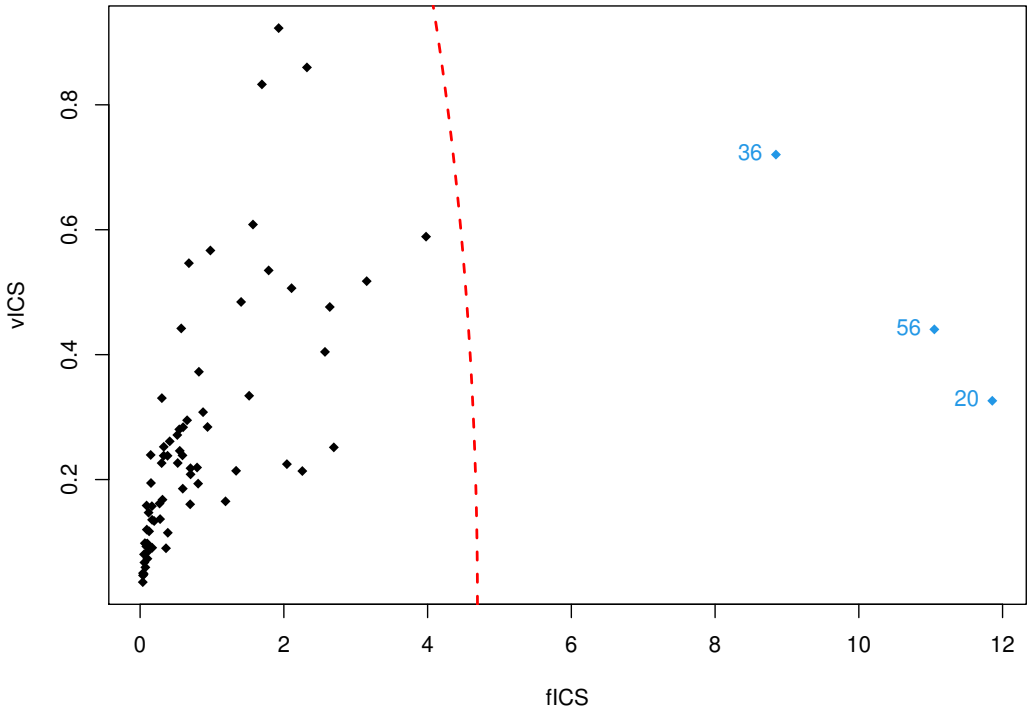

Figure 4.2-2 – **Weather data set** - Point-wise ICS: FOM with the automatic dimension selection and the cutoff quantile of order 0.95.

where med denotes the median. An observation i is flagged as an outlier if

$$\frac{\text{LICS}_i - \text{med}(\text{LICS})}{\text{MAD}(\text{LICS})} > \Phi^{-1}(\alpha), \tag{4.2.4}$$

where MAD denotes the median absolute deviation, Φ the standard normal cumulative distribution function and α a quantile order. For $\alpha = 0.95$, expression (4.2.4) yields the dashed red curve on Figure 4.2-2 which is part of an ellipse. The scatterplot clearly distinguishes the curves 20, 36 and 56 from the other curves both in terms of fICS and vICS (with slightly more variability for observation 36 than for observations 20 and 56). As already noticed in Figure 4.2-1, these three curves have a large average but also have widely dispersed ICS distances over time.

The impact of the dimension reduction step can be analysed by looking at Figure 4.2-3, where we give the FOM when the number of selected components is set to $k = 1$ (resp., $k = 2$ and $k = 3$) during the whole time period, in the left (resp., middle and right) panel. As anticipated, the number of observations flagged as outliers increases when the dimension increases. Note that the case $k = p = 3$ corresponds to the Mahalanobis distance and leads to the detection of most of the 17 curves detailed previously and suspected of atypical behaviour (in particular, the green and red curves in Figure 4.6-25). In contrast, selecting $k = 1$ gives a similar result to the result obtained

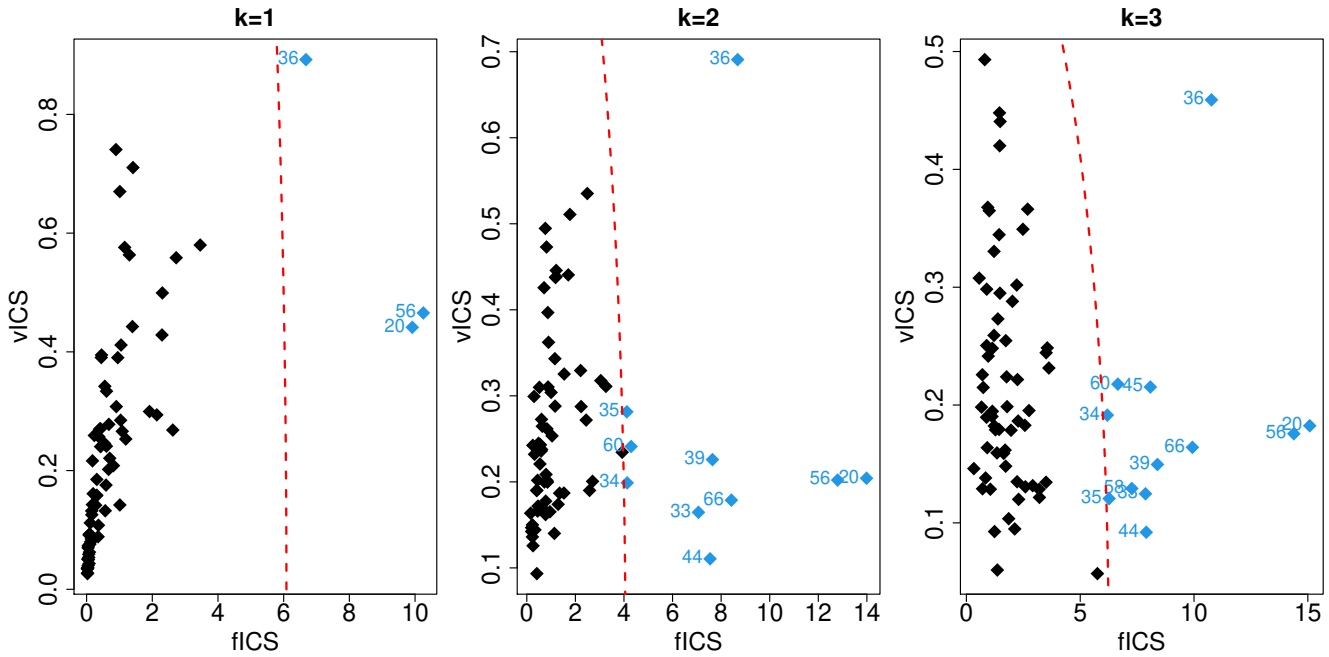

Figure 4.2-3 – **Weather data set** - Point-wise ICS: FOM with $k = 1$ (resp., $k = 2$ and $k = 3$) on the left (resp., middle and right) panel and the cutoff quantile of order 0.95.

with automatic selection in Figure 4.2-2. The result obtained with $k = 2$ is in between, with essentially the curves 33, 39, 44 and 66 (the green curves on Figure 4.6-25) detected in addition to the curves 20, 36 and 56. In the context that we are interested in, which has a small proportion of outliers to detect, we prefer the automated selection or the choice $k = 1$. The question that arises now is how can we interpret the outlyingness of the three observations 20, 36 and 56.

Interpreting outlying curves in the context of point-wise ICS is possible but not easy because of the temporal dimension combined with the possible presence of many variables. For the weather example, it is possible to interpret the three outlying curves 20, 36 and 56 by looking closely at Figures 4.6-23 and 4.6-24. Curve 36 is unique in the sense that it has a very particular wind speed curve with very high values and a unimodal and peacked shape at approximately the 200th day. Observation 56 is also very special in the sense that it takes quite large values in terms of the wind speed, with some small bumps at approximately the 50th and 300th days and a clear decrease at approximately the 200th day. Other curves exhibit large values for the wind speed and, in particular, all of the red curves in Figure 4.6-23 have large and unusual behaviour in terms of the temperature. However, observation 56 is the only curve among the red curves that has large wind speed values together with small temperature values. Observation 20 is even more atypical than 56 in terms of having large wind speed values, even though its temperature values are not very small. Note that curve 59 is also quite different from all of the other curves that have large values of wind speed and medium temperature values. However, as seen in Figure 4.2-4, curve 59 (in cyan) is not detected using point-wise ICS.

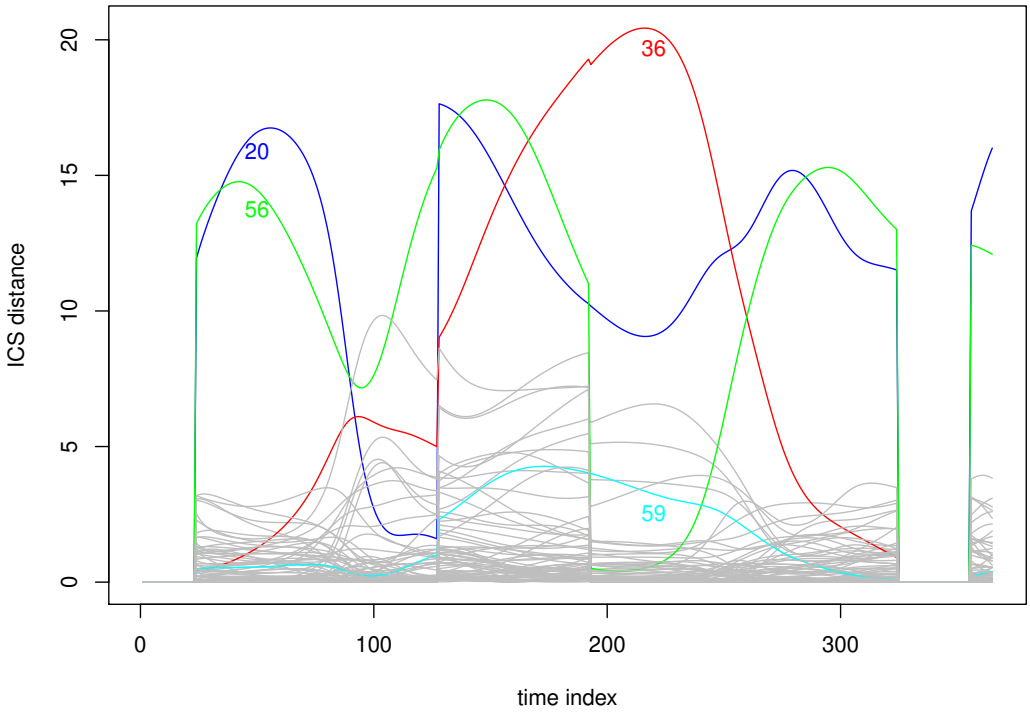

Figure 4.2-4 – **Weather data set** - Point-wise ICS: square ICS distance.

To facilitate the interpretation, we propose to calculate at each time point the correlations between the selected invariant components and the initial variables. Examining such correlation curves is possible if the number of initial variables and the number of selected components are not too large, which is the case for the weather data set, where $p = 3$ and the number of selected components is not larger than 2. The plot on the left (resp., middle and right) panel of Figure 4.2-5 gives the correlation curve between the temperature (resp., wind speed and log precipitation) and the first ICS component. Values larger than 0.20 in absolute value are plotted in red. Note that the range of values differs from one plot to another, to zoom in. It is clear that the first invariant component is highly positively correlated with the wind speed with a smaller correlation between the 100th and 200th days. During the same period of time, the correlation of the first component becomes negative with the log precipitation and moderately negative with the temperature. These plots in conjunction with Figure 4.2-4 confirm our previous interpretation of the outlying curves 20, 36 and 56. At approximately days 50, 200 and 300, the first component is strongly positively correlated with the wind speed, and observations 20 and 56 (resp., 36) take very large wind speed values at days 50 and 300 (resp., 200). Curves 20 and 56 are also outlying at day 150, when the component is positively correlated with the wind speed but also negatively correlated with the temperature and the log precipitation, which corresponds to the fact that these curves have large wind speed together with a small temperature and log precipitation. We do not give the plot of

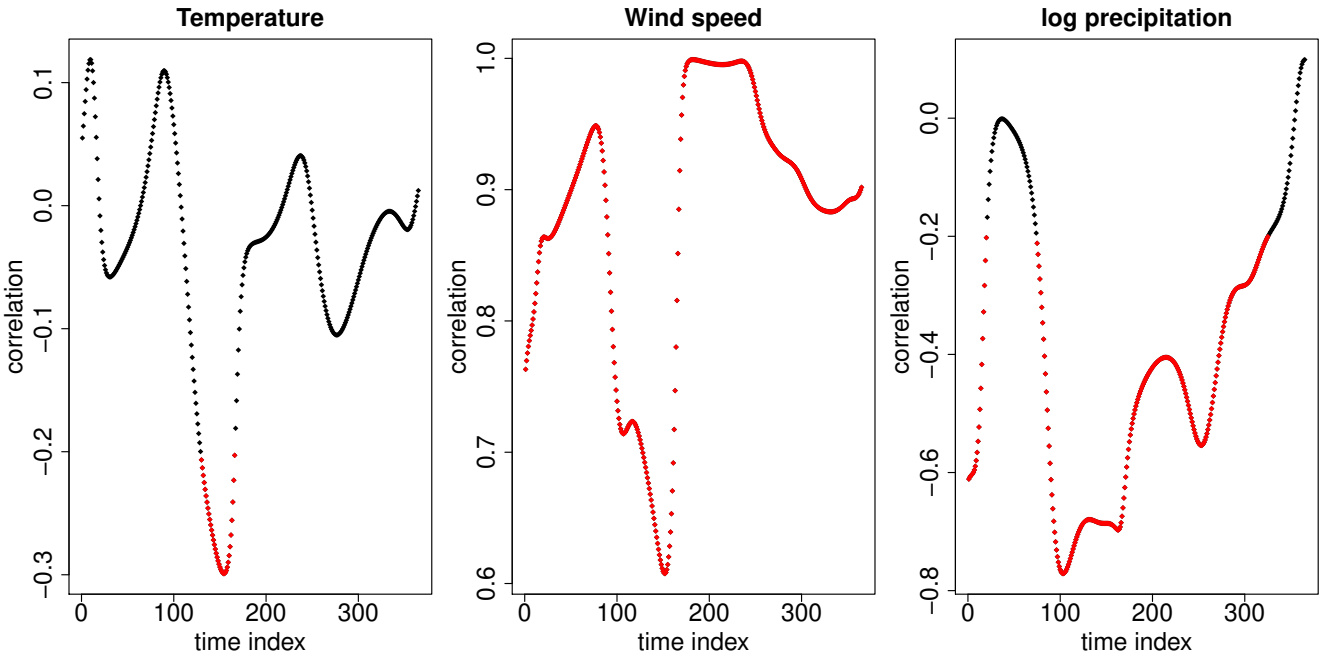

Figure 4.2-5 – **Weather data set** - Point-wise ICS: correlations between the first component and the three initial variables by time. Red colour corresponds to correlation with absolute value larger than 0.20.

the correlations for the second component, but this component is highly positively correlated with the wind speed only between day 100 and day 200. This finding explains the fact that curve 36 is closer in terms of the ICS distance from 20 and 56 in Figure 4.2-2, where the second component of ICS accounted for between day 120 and day 180 (see the left plot of Figure 4.2-1) compared with Figure 4.2-3, which has only one component (left plot). The second component is also very much positively correlated with the log precipitation during the first 100 days which explains that the green curves in Figure 4.6-25 are detected as outliers on Figure 4.2-3 when $k = 2$ at every time (middle plot).

In the case of large p, the previous interpretations could become intractable, and a global ICS, as detailed below, could be a good alternative.

4.2.4 Global functional ICS

For each functional output variable, the curves are aligned and projected onto a truncated functional basis with D dimensions. The data set is thus made of n observations and pD variables given by the coefficients along the basis. As long as $n > pD$, we apply an ordinary ICS and obtain pD invariant components.

We could then follow Archimbaud et al. (2018) and use a test procedure to select the number of invariant components as the test proposed by Nordhausen et al. (2017) and detailed previously.

We instead use the scree plot, as advised by Archimbaud et al. (2018), and spot an elbow in the decrease of the eigenvalues. This graphical method is simple and tends to select fewer variables than the test procedure which selects the eigenvalues significantly larger than one. It cannot be used in the context of point-wise ICS, where the dimension selection must be performed at each time point and must be automated, but we recommend its use for global ICS. Figure 4.2-6 gives the scree plot of the global ICS for the weather data set. There are $p \times D = 3 \times 11 = 33$ eigenvalues and we have several possible dimension choices. We decide to choose two (squares on the plot), three (squares and triangle) or four (squares, triangle and cross) components and compare the results.

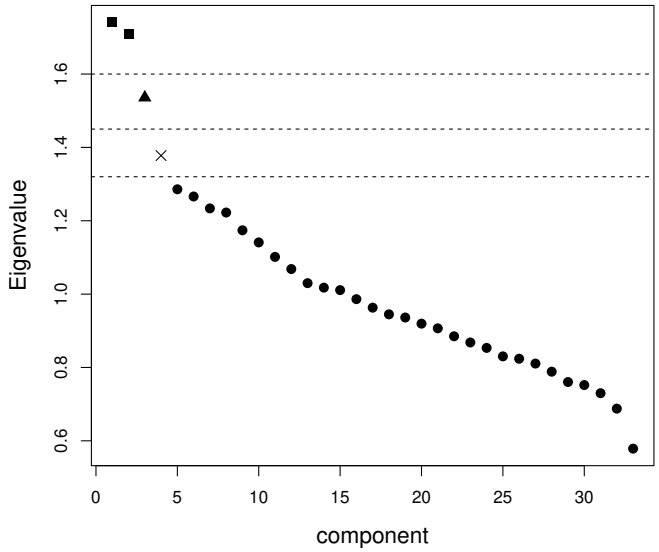

Figure 4.2-6 – **Weather data set** - Global ICS: scree plot.

We then calculate and plot the ICS distance of each observation using the expression (4.2.1) for the given number of selected components. We also plot a cutoff line obtained by Monte Carlo simulations as proposed by Archimbaud et al. (2018) and previously detailed. Figure 4.2-7 gives the ICS distance plot for $k = 2$ (resp., $k = 3$ and $k = 4$) selected components on the left (resp., middle and right) panel. Selecting only two components leads to detecting curves 20 and 56 as outliers while selecting three (resp., four) components leads to also detecting curve 59 (resp., 59 and 36).

It is possible to interpret the outliers by looking at the correlations between the basis coefficients of each initial variable and the selected invariant coordinates. Figure 4.2-8 gives the correlations for each of the four components in different plots. On every plot the points correspond to the different B-spline coefficients (11 per variable) and they are grouped by variable in 3 columns, with the temperature on the left side of the first vertical green line, the wind speed in between the first and second green lines, and the log precipitation on the right of the second green line. Interestingly, we

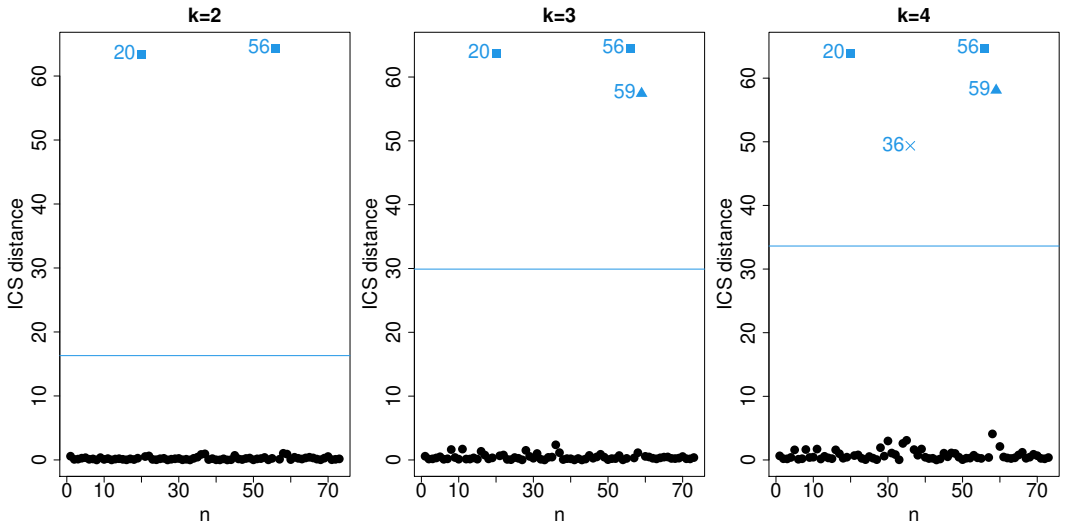

Figure 4.2-7 – **Weather data set** - Global ICS: square ICS distance with $k = 2$ (resp., $k = 3$ and $k = 4$) selected components on the left (resp., center and right) panel and the cutoff quantile of order 0.975.

note that the correlations are quite similar for a given component among the different coefficients of the same variable. For the first three components which lead to the detection of curves 20, 56 and 59, the correlation structure is globally similar and has negative correlations between the components and wind speed coefficients and low or medium positive correlations with the temperature and log precipitation coefficients. The first three plots differ when looking at the correlations with the different B-splines coefficients in detail. However it is not easy to interpret these coefficients. With regard to the correlations with the fourth component on the right bottom plot, they differ from the other plots and exhibit positive correlations with both the temperature and the wind speed, and negative correlations with the log precipitation. This finding explains that curve 36 which has a very high level of wind speed together with a large temperature is detected with $k = 4$. Note that Fig. 9 from the Dai et al. (2020) plot, for the same weather data set, give joint outlying curves that are not detected using univariate methods, and among them are observations 56 (green on Fig. 9) and 59 (purple on Fig. 9).

Each of the two functional ICS approaches has advantages and disadvantages. One disadvantage of the point-wise approach is that it does not account for the temporal dependence in the data. This limitation is not the case for the global approach which accounts for the temporal behaviour of the data through functional dimension reduction. However, global ICS depends on the choice of a functional basis, and interpreting the outlying curves using the coefficients in the basis is not easy to accomplish. An advantage of global ICS is that the domains for the functional variables do not need to be the same. Because it only involves one joint diagonalization, global ICS is much less expensive in terms of the calculation time than point-wise ICS. Additionally, it uses one and only one dimension selection, which can be made using the scree plot. With regard to point-wise

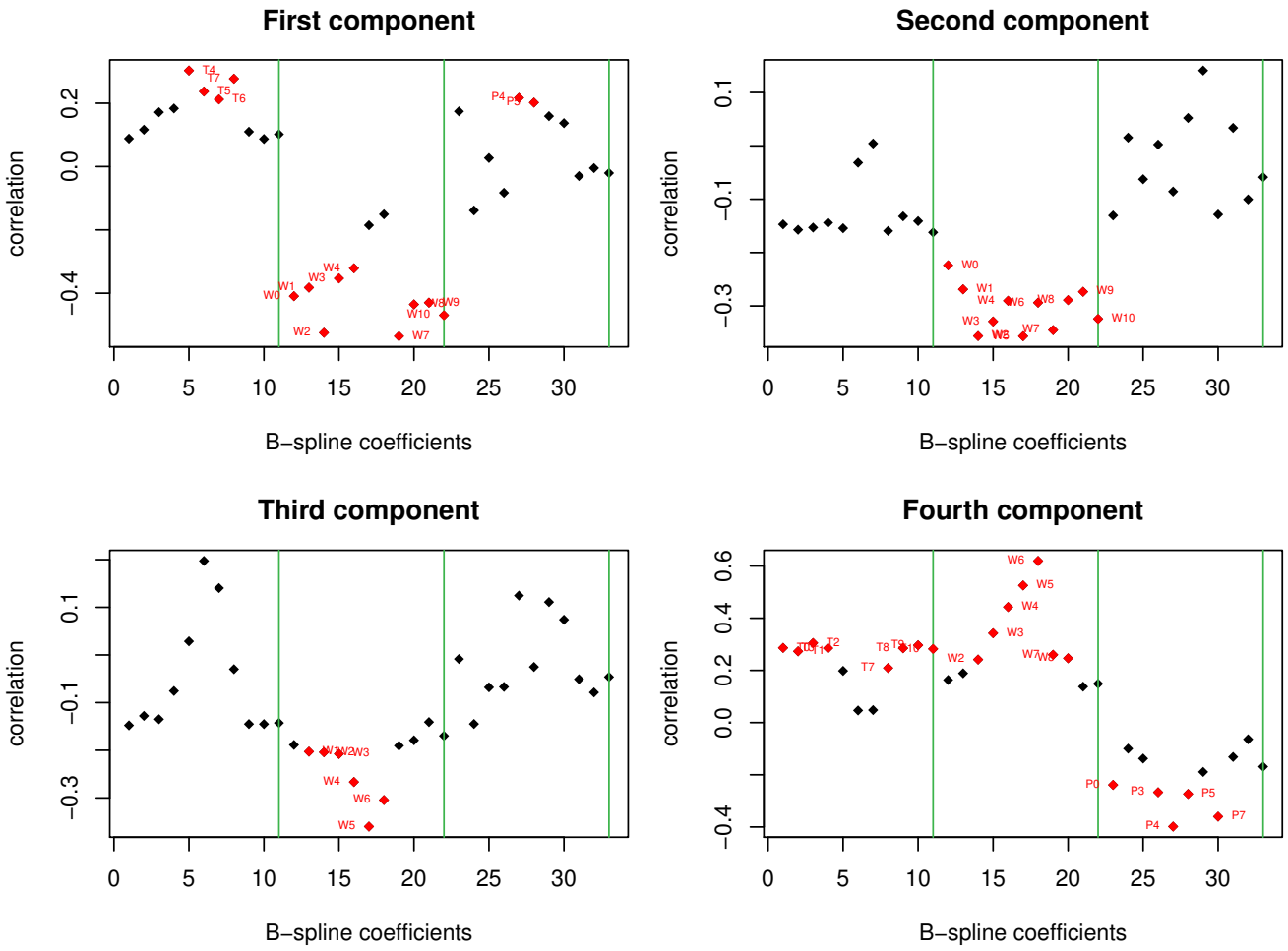

Figure 4.2-8 – **Weather data set** - Global ICS: correlations between the first four components and the 11 B-spline coefficients of the three initial variables. Red colour corresponds to correlation with absolute value larger than 0.20.

ICS, it allows for a graphical representation of the results in the form of curves, which is helpful in deepening the results.

4.3 Data analysis

We consider two real data sets that come from the fields of aeronautics and microelectronics manufacturing. The first example concerns predictive maintenance of an aeronautical electrical generator before its failure. The second example concerns a quality control process in a semiconductor microelectronics fabrication. The first data set is completely confidential, while the second is available on the web.

For each of these two data sets, we have information that concerns the observations that are

considered to be true outliers. Thus, we can evaluate the performance of the approaches by counting the number of true and false positives.

Moreover, for both examples, we consider that only a small proportion of the observations is expected to be diagnosed as anomalies, and an effective method must allow for the detection of the true positives with as few false positives as possible.

The goal is to compare the two functional ICS approaches to detecting outlying curves. We also include in the comparison the directional outlier detection method proposed by Rousseeuw et al. (2018) using a point-wise Directional Outlyingness (DO) measure. Both point-wise ICS and the method by Rousseeuw et al. (2018) make use of the functional outlier map (FOM) tool. We use the *fOutl* and the *fom* functions available in the R package *mrfDepth* to implement Rousseeuw et al. (2018). The *fom* function is customized to be able to change the colours and the cutoff quantile order. In the absence of auxiliary information, we use the default uniform values for the weights.

For the function expansion used in global ICS, we consider the Fourier basis but note that the results obtained with a B-splines basis are similar. Concerning the choice of the dimension D, based on our experience in using global ICS on many real examples, the number of observations n and the number of features p are to be taken into account. The number of observations per dimension $p \times D$ should not be smaller than 10. This criterion gives the rule of thumb that D should be smaller than $n/(10p)$.

In the main body of the text, we focus on a particular D value in $\{5, 11, 15\}$ using the recommended rule of thumb. We compare the results with the other D values for the quality control example in the appendix.

4.3.1 Predictive maintenance of an aeronautical electrical generator

This first application is an example of aircraft flight data. In the aeronautics literature, several recent papers tackle the problem of anomaly detection on flight data even if the literature is still rather sparse. Most articles consider a multivariate time series framework (see Li et al. (2015), Li et al. (2016), Memarzadeh et al. (2020) and the references therein). One exception is the paper by Jarry et al. (2020) which considers PCA in a univariate functional framework and a clustering method, to detect atypical approaches using landing radar records.

Our application concerns the monitoring of aircraft electrical generators during routine flights. Electrical generator failures lead to delays or cancellations of flights which can be extremely costly to the manufacturer and airline. To reduce this cost, electrical design engineers are willing to detect a generator abnormal behaviour before it turns into a failure. The goal is to raise a warning when a sequence of abnormal flights is detected and to suggest performing a maintenance action on the generator. Such a process is called predictive maintenance. We consider $n = 590$ flights of

Table 4.3.1 – Recorded features by flight to detect electrical generator abnormal behaviour.

Notation	Description	Unit
X^1	Generator oil temperature	C°
X^2	Engine speed	Knots (kts)
X^3	Generator load	KVA
X^4	Static air temperature	C°
X^5	Computed air speed	kts
X^6	Altitude	ft

distinct duration from a given aircraft and a given generator. For each flight, we observe $p = 6$ features with a sampling rate of one record per second. These features are identified as relevant by electrical engineers and are detailed in Table 4.3.1. To be able to apply the selected approaches we must align the flights to obtain the same number T of time points. To account for the different flight phases, which correspond to different electrical behaviours, we split each flight into takeoff, cruise and landing phases. Each flight phase is aligned separately, and the whole flights are rebuilt afterward. In the present example, the flights are aligned on the duration of $T = 2900$ seconds.

In Figure 4.6-26 of the appendix, we plot the aligned flights by the features for the engine speed, the static air temperature, the computed air speed and the altitude. For confidentiality reasons, we cannot plot the generator oil temperature and the generator load curves but we use the data in the analysis. The four flights that precede the generator loss are considered to be abnormal flights and are coloured in red. Given that the size of the sample is $n = 590$, the percentage of outliers being between 1 and 2% corresponds to the number of detected outliers being between 6 and 12. Let us now implement and compare the two functional ICS approaches on this real data set. Note that looking at Figure 4.6-26, the four red curves are clearly outlying for the engine speed feature during the landing period.

First, we apply global ICS on the aligned flights using $D = 5$ since the rule of thumb $D < n/(10p)$ leads to $D < 9$.

With $D = 5$, we have a data set with dimension 590×30. Global ICS consists in computing the square distances using the first k invariant components. To select k, we use the scree plot (see Figure 4.3-9), and we identify the three possible values for k, $k = 1$ (square symbol), $k = 2$ (triangle) and $k = 4$ (diamond).

For each k value, we compute the squared ICS distances by flight and the ICS cutoff using the Monte Carlo calculation with 100 replications and a level $\gamma = 0.025$ (function *dist.simu.test* from the R package *ICSOutlier*).

The results are given in Figure 4.3-10, where the flights are ordered by date, and we plot a black dashed vertical line that represents the flight that precedes the generator loss. The symbols

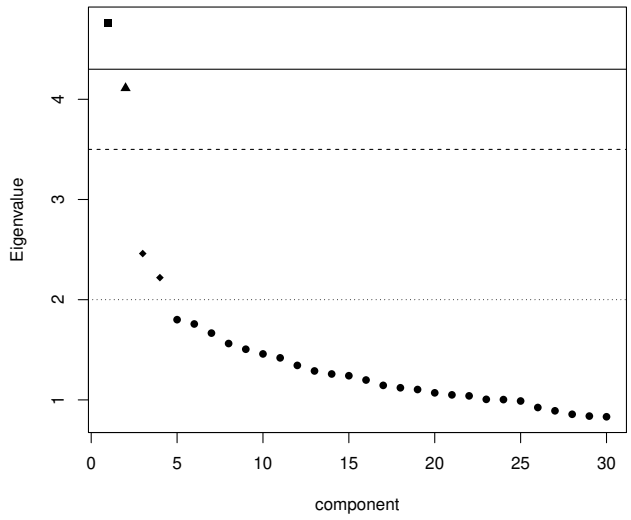

Figure 4.3-9 – **Aeronautical data set** - Global ICS: scree plot.

of the flights that are detected as abnormal are the same as the symbols used in the scree plot (square when detected on the first component, triangle on the second and diamond on the third and fourth components). The true positive flights are coloured in red, and the false positive flights in blue. With $k = 4$, we have already too many outliers compared to the 2% limit and thus we should consider $k = 1$ or $k = 2$. When $k = 1$, only the true outliers are detected which is the ideal situation. However, even with $k = 2$, the only successive flights detected are the four that precede the generator loss. Looking at the correlations of the first invariant component with the 5 Fourier coefficients of the engine speed feature in Figure 4.6-27, we can see that the four flights differ from the others because of the engine speed variable, as we have already noticed.

Then, we apply the point-wise functional ICS on the aligned curves and use the automatic components selection procedure described previously, which follows Nordhausen et al. (2017). The number of selected components by time is given in Figure 4.3-11. This number varies from two to six and is higher during the cruise period than during take-off and landing. Given the high variability of the number of selected components, we choose to divide the square ICS distances by the number of components at each time point when calculating fICS and vICS. Using fICS and vICS, the FOM is plotted on the left panel of Figure 4.3-12. The cutoff curve is calculated with the quantile of order 0.999995. This value has been adjusted in order not to detect too many outliers. The detected outliers are coloured in red (the true ones) or in blue (the additional ones). The central panel of Figure 4.3-12 shows the binary outlyingness indicator by flight while the right panel gives the fICS values. It can be seen that the true outliers are detected together with six other flights (11, 17, 34, 53, 308 and 456), and among them, 4 flights (11, 17, 53 and 456) are also detected by global ICS with $k = 2$ (see the middle panel of Figure 4.3-10). The

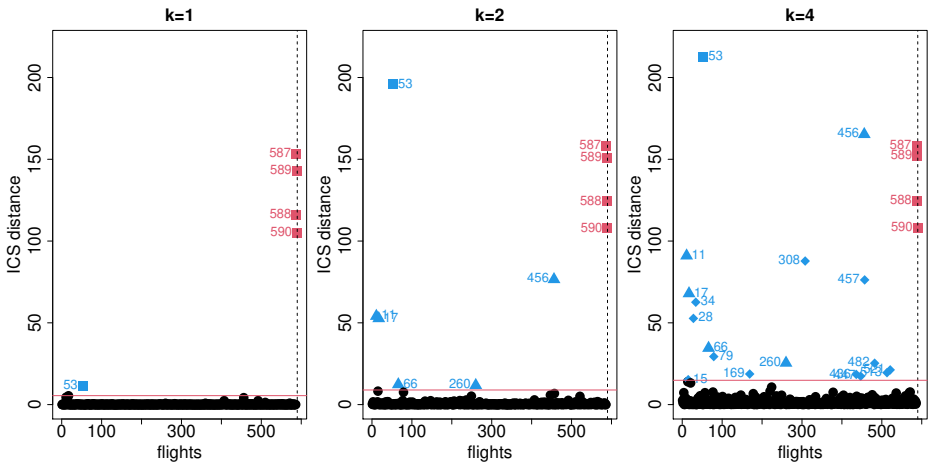

Figure 4.3-10 – **Aeronautical data set** - Global ICS: square ICS distance with $k = 1$ (resp., $k = 2$ and $k = 4$) selected components on the left (resp., center and right) panel and the cutoff quantile of order 0.975.

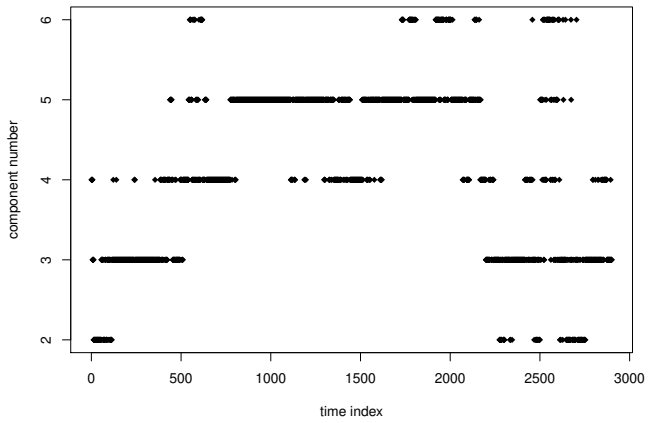

Figure 4.3-11 – **Aeronautical data set** - Point-wise ICS: number of selected components.

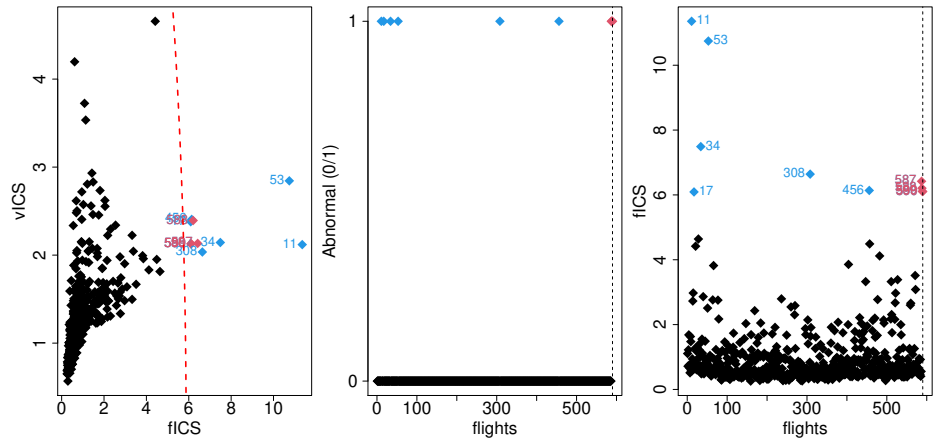

Figure 4.3-12 – **Aeronautical data set** - Point-wise ICS (with square ICS distances by flight divided by the number of automatically selected components at each time point), Left: FOM with the cutoff quantile of order 0.999995, Center: outlier flag (1 for abnormal and 0 for normal flights), Right: fICS.

square ICS distances are plotted in Figure 4.3-13 with the same colour code as in Figure 4.3-12 (red for the true outliers and blue for the additional outliers detected by point-wise ICS). It can be seen that the red curves differ from the other curves at the end of the flights while the blue curves are outlying during the cruise period. It is however, difficult to analyse the reasons for the outlyingness of the red and blue flights by looking at the correlation curves because of the high number of initial features and the large and highly variable number of components.

To visualize the flights detected as anomalies by global and point-wise ICS and to interpret their outlyingness, we plot the original curves with different colours in Figure 4.3-14 (see the caption for the details on the colours). It can be seen that all of the flights that are detected as outliers either by global or point-wise ICS exhibit a different behaviour compared to the other flights. Global and point-wise ICS do not give exactly the same results but have many curves detected as outliers in common.

We now compare the ICS results with the FOM obtained using the DO outlyingness index as defined in Rousseeuw et al. (2018). Using the default options leads to very bad results (see Figure 4.6-31); as a result, we tried out other options (available but not detailed in the documentation) that lead to Figure 4.3-15, which permits us to detect all of the true positive outliers with no false positives. The method is, however, quite unsatisfactory because for almost half of the time points (46%), an exact fit is detected and it is not possible to calculate the outlyingness index. Figure 4.6-32 in the appendix gives the time points at which it is not possible to make the calculus of the DO index and they correspond to the flights cruise. This finding could explain why the method does not detect any of the curves 11, 17, 34, 53, 66, 260, 308 and 456 which are outlying during the cruise period.

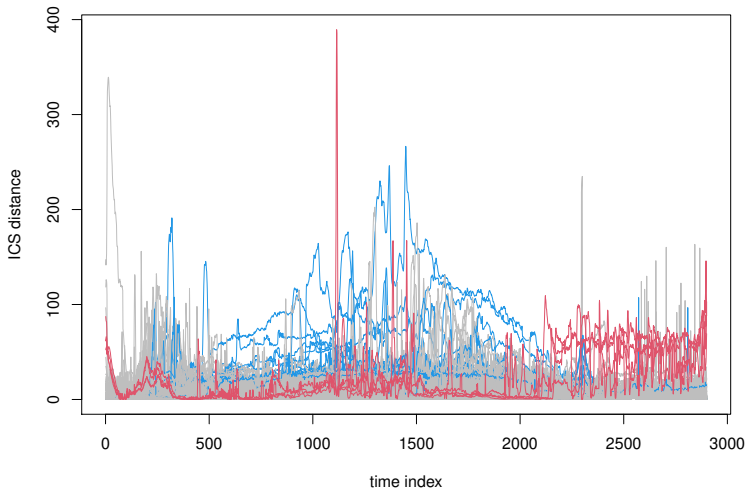

Figure 4.3-13 – **Aeronautical data set** - Point-wise ICS: square ICS distances with an automatic components selection. Normal flights are coloured in grey whereas the flights that precede the failure are coloured in red and the additional flights detected as outliers are in blue.

4.3.2 Semiconductor quality control application

Anomaly detection is crucial in quality control and especially in semiconductor microelectronics, which has safety-critical applications, such as automotive electronics, medical devices, and aerospace systems. Archimbaud (2018a) gives a review of the common unsupervised methods that are used in practice together with their implementation in R software. It appears that only a few multivariate methods, such as Mahalanobis distance or Principal Components Analysis, are used by manufacturers. Some recently published articles in the field of industrial process monitoring consider a multivariate functional framework (see for example Liu et al. (2020) and the references therein).

The database that we consider is available online[1]. It contains a collection of sequences of measurements (or runs) that are recorded by one vacuum-chamber sensor during the etch process applied to one silicon wafer during the manufacture of semiconductor microelectronics. An etch process is a complex process during which layers of various materials are applied to a silicon wafer and selectively removed to define circuit elements on the wafer. Each run, among the 1194 runs, has an assigned classification of normal or abnormal and 127 runs are flagged as abnormal. This number corresponds to 10.6% of the observations which is much larger than the 2% we are interested in. Thus, we only consider as true outliers the four runs with severe or very severe faults which correspond to 0.3% of the observations. Six sensors have been identified by domain experts

[1]Website: `https://www.cs.cmu.edu/~bobski/data/data.html`.

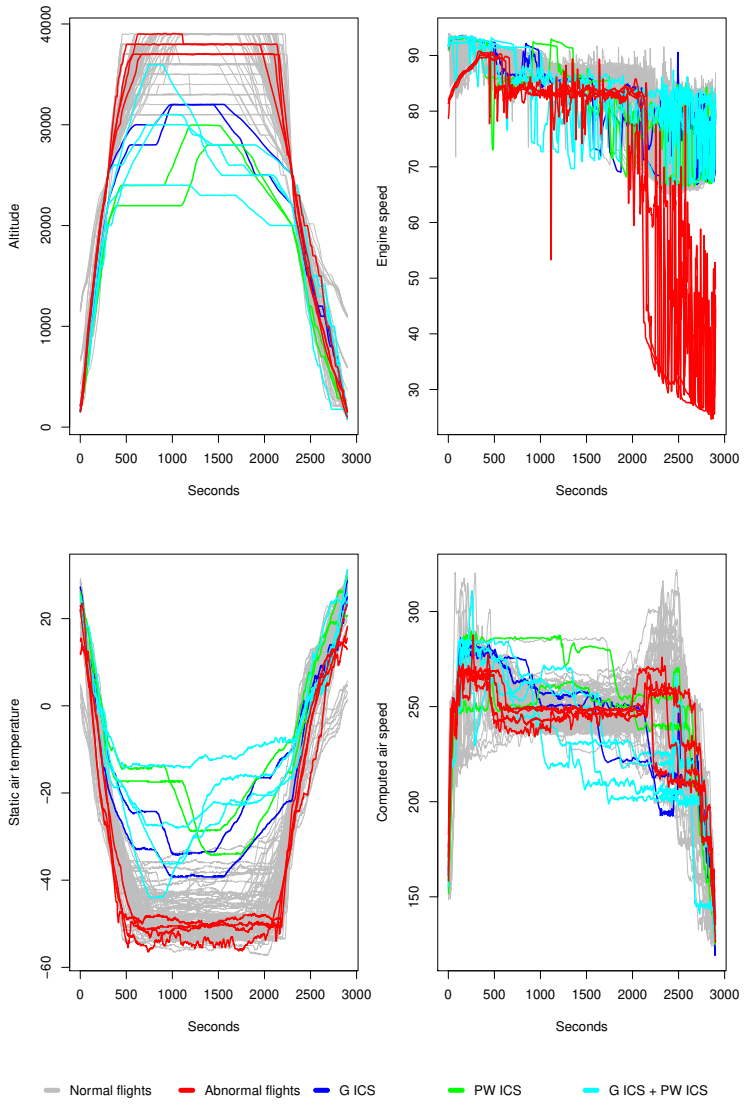

Figure 4.3-14 – **Aeronautical data set** - Detected outliers. Normal flights are coloured in grey whereas the flights that precede the failure are coloured in red. The curves 66 and 260 detected as outliers only by global ICS with $k = 2$ (G ICS) are in blue, the curves 34 and 308 detected as outliers only by point-wise ICS (PW ICS) with an automatic components selection are in green, while the curves 11, 17, 53 and 456 detected as outliers by G ICS and PW ICS are in cyan.

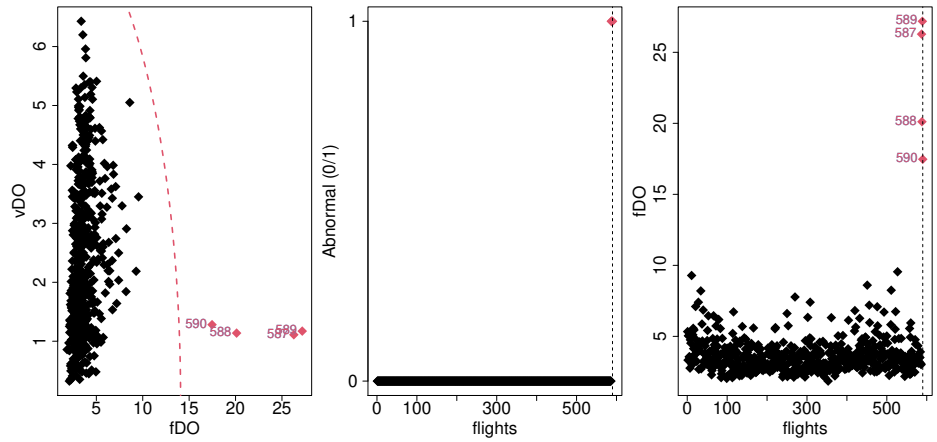

Figure 4.3-15 – **Aeronautical data set** - Directional Outlyingness method using the R package mrfDepth (with distOptions = list(rmZeroes = TRUE,maxRatio = 3) in the fOutl function). Left: FOM with the default quantile of order 0.995, Center: outlier flag (1 for abnormal and 0 for normal flights), Right: fDO by flight.

Table 4.3.2 – Sensors description for the semiconductor manufacturing data set.

Notation	**Sensors description**
Sensor 6	radio frequency forward power sensor
Sensor 7	radio frequency reflected power sensor
Sensor 8	chamber pressure sensor
Sensor 11	405 nm emission sensor
Sensor 12	520 nm emission sensor
Sensor 15	direct current bias sensor

as being critical for monitoring purposes (see Olszewski (2001), p. 68-69, for more details on the sensors). The description of the $p = 6$ sensors is given in Table 4.3.2. We have $n = 1194$ runs with a duration that varies from 104 to 198 records. The runs are aligned by linear interpolation to a duration of $T = 150$ records. The runs after alignment are plotted in Figure 4.6-33 for each sensor. The four abnormal runs 73, 107, 317 and 351 are highlighted in red.

For global ICS, a dimension reduction is applied on the aligned runs for each sensor using the Fourier basis, with the number of basis functions equal to $D = 11$. The appendix gives the results for other values of D but applying the rule of thumb $D < n/(10p)$ leads to $D < 19$, and thus, we focus on $D = 11$. The global ICS results on this 1194×66 data set are given on Figures 4.3-16 and 4.3-17. The scree plot on Figure 4.3-16 shows that we can consider $k = 1$, 3, 6 or 7 and we use different symbols to differentiate between the eigenvalues (square for the first two eigenvalues, triangle for the third, diamond for the fourth to the sixth and cross for the seventh). Figure 4.3-17 gives the square ICS distances and the ICS cutoff (red line) for each of the four k value. To

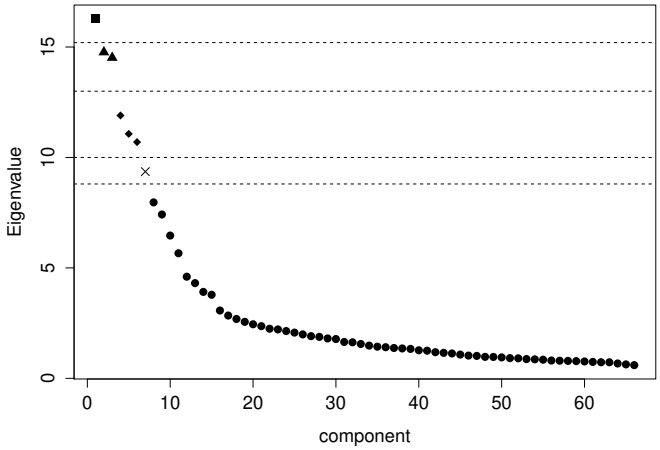

Figure 4.3-16 – **Semiconductor data set** - Global ICS: scree plot.

compute the ICS cutoffs we use again the function *dist.simu.test* from the R package *ICSOutlier* with a quantile of order $1 - \gamma = 0.975$. We coloured the true positive runs in red whereas the false positive runs are in blue and the symbols correspond to the symbols in Figure 4.3-16. Using $k = 6$ or 7 leads to the detection of a dozen observations which corresponds to approximately 1% of the observations. Such a choice leads to the detection of the four outliers together with some other runs.

For point-wise functional ICS, we apply $T = 150$ ordinary ICS with $n = 590$ and $p = 6$. For each ordinary ICS, we select the number of invariant components automatically as detailed in Section 4.2.3 and the plot is given in Figure 4.3-18. The plot is not well structured and has large amount of variability in the number of selected dimensions. Using the resulting ICS distances, we give three different FOM: (i) with an fICS weighted by the number of components selected at each time in Figure 4.3-19, (ii) with an unweighted fICS in Figure 4.3-20 and (iii) with a fixed number of selected components $k = 1$ in Figure 4.3-21. Once again, it can be seen that the choice of the dimension is crucial. If we tolerate a detection rate of 1%, we can use the weighted fICS and detect the four true outliers with six false positives. We note that the runs 122, 964, 311 and 326 are detected by both functional ICSs but that some more observations are detected by only one of the methods. Note also that if we except the run 162 for global ICS and 489 for point-wise ICS which are normal runs, the runs declared as outliers are abnormal runs (but with no severe or very severe fault).

We also give the FOM for the Directional Outlyingness index proposed by Rousseeuw et al. (2018) in Figure 4.3-22.

Once again, some exact fit problems do not allow us to calculate the DO index 50% of the time (see Figure 4.6-36 for a plot of the time points, where the calculus is not possible). This finding could explain that only two observations are detected as outliers with one true positive among

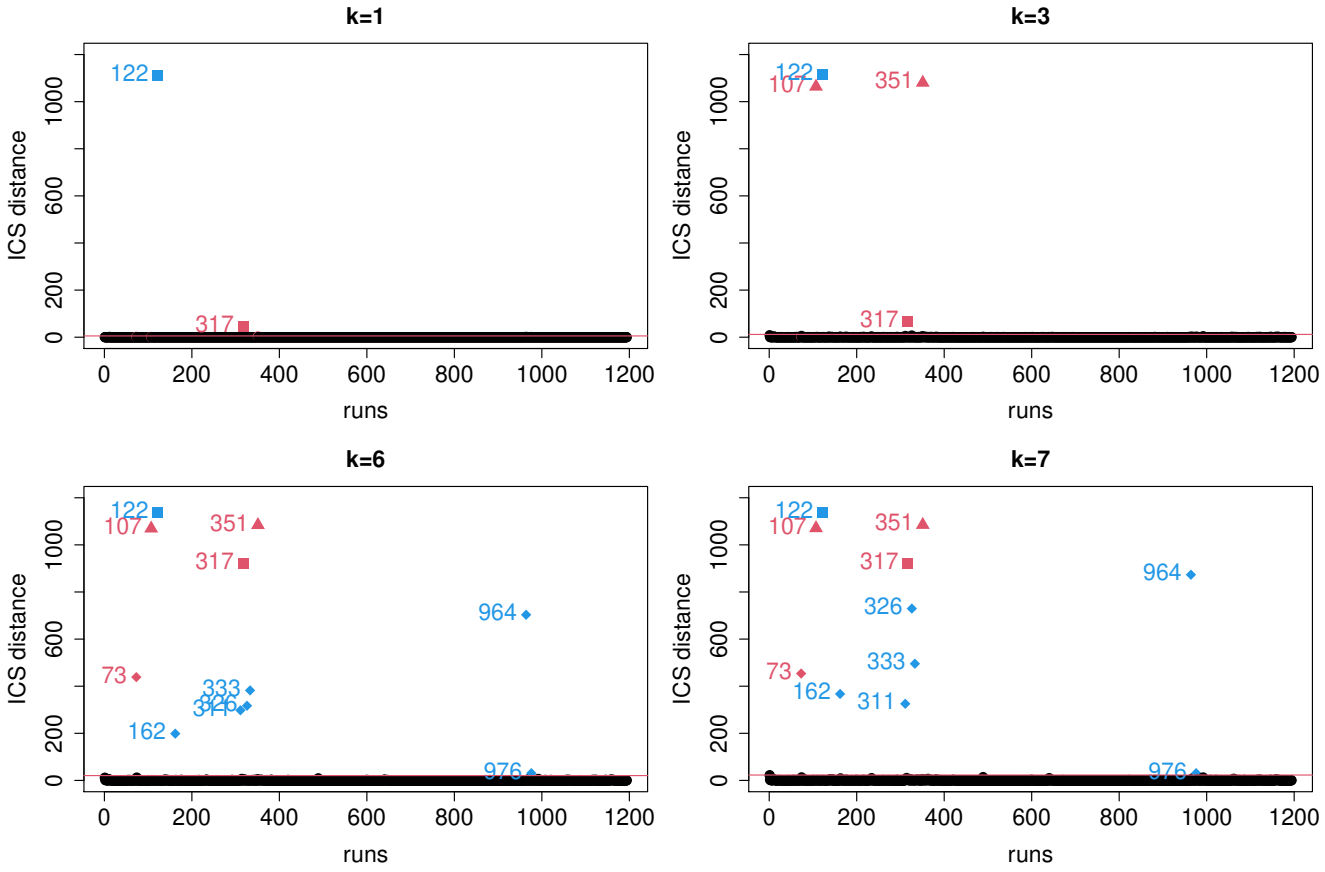

Figure 4.3-17 – **Semiconductor data set** - Global ICS: square ICS distance for $k = 1$, 3, 6 and 7 using the cutoff quantile of order 0.975.

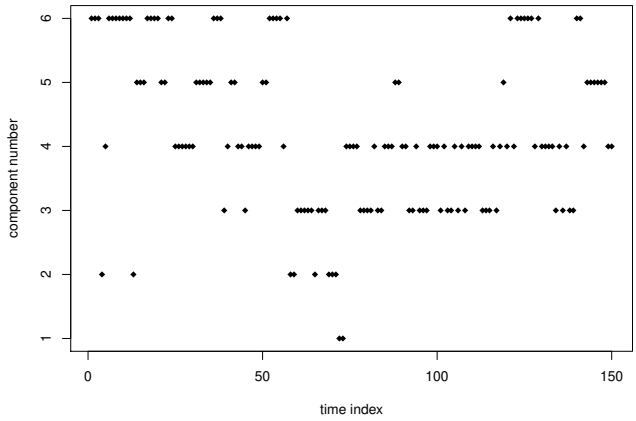

Figure 4.3-18 – **Semiconductor data set** - Point-wise ICS: Number of selected components.

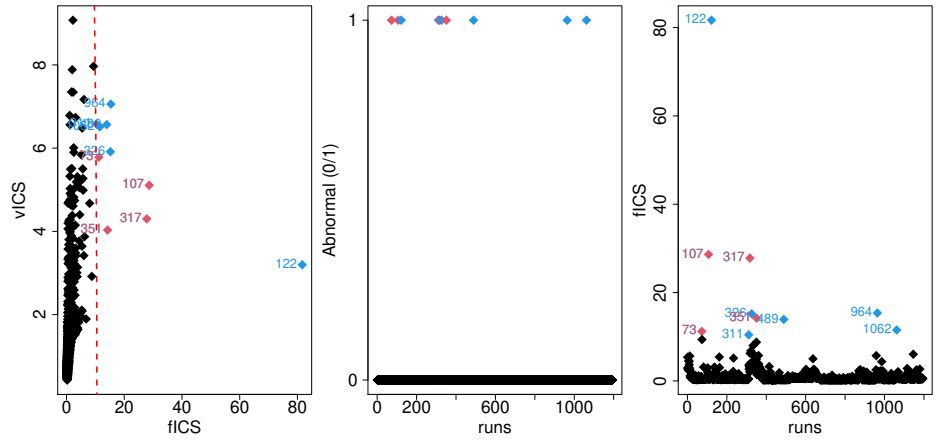

Figure 4.3-19 – **Semiconductor data set** - Point-wise ICS (with square ICS distances by run divided by the number of automatically selected components at each time point), Left: FOM using the cutoff quantile of order 0.99999995, Center: outlier flag (1 for abnormal and 0 for normal flights), Right: fICS.

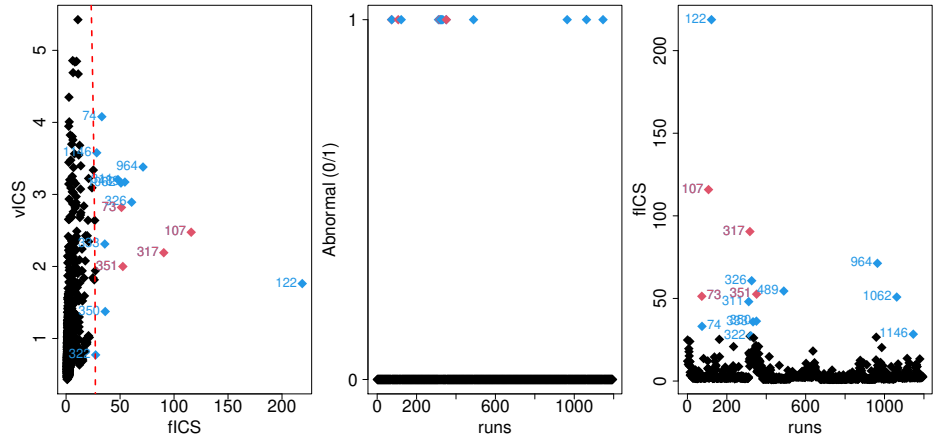

Figure 4.3-20 – **Semiconductor data set** - Point-wise ICS (with square ICS distances by run at each time point), Left: FOM using a cutoff quantile of 0.99999995, Center: outlier flag (1 for abnormal and 0 for normal flights), Right: fICS by run.

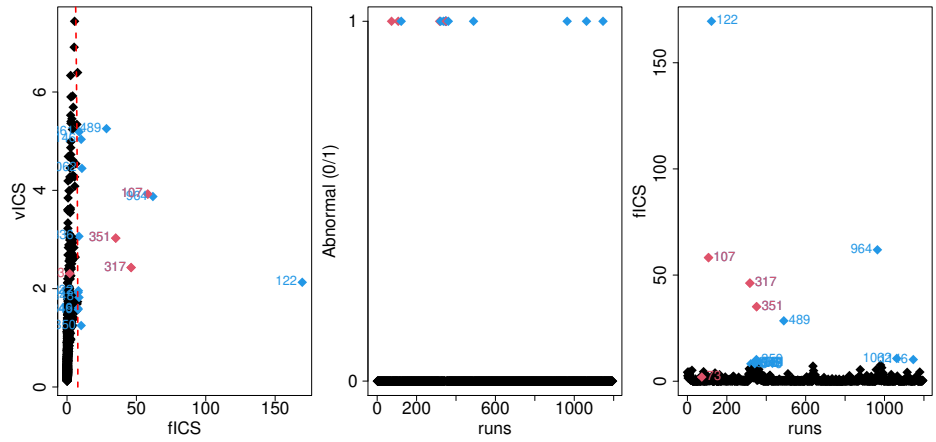

Figure 4.3-21 – **Semiconductor data set** - Point-wise ICS (with $k = 1$), Left: FOM using the cutoff quantile of order 0.99999995, Center: outlier flag (1 for abnormal and 0 for normal flights), Right: fICS by run.

them. Changing the options of the R functions in this example does not change the results (see Figure 4.6-37 in the appendix).

4.4 Conclusions and perspectives

The present paper propose two generalizations of ICS to functional data with many diagnostic plots to help the data analyst in the detection and interpretation of outlying curves in a multivariate framework. Both approaches have advantages and disadvantages and can be used in a complementary manner. Given the complexity of outlier detection in the framework of multivariate functional data, it appears to be unrealistic to find a method that works in all situations and the methodology that we propose is particularly suitable when there is a small number of observations that are likely to be real anomalies that cannot be identified when looking at univariate functional characteristics only. In the context of a small proportion of outliers, it is important to have detection methods that lead to the identification of only a few anomalies with the possibility to understand their anomaly behaviour. Functional ICS, either global or point-wise, with the covariance matrix and the matrix of fourth moment as the scatter pair, is particularly suitable for such a context.

Among the research perspectives for the global ICS method, there is the possibility of making vary the choice of the functional basis and the number of coefficients in the basis expansion, according to the feature, or the possibility of taking into account the Gram matrix for non-orthonormal bases as suggested by Virta et al. (2020). Let us also mention the case of curves measured on different domains (Happ and Greven, 2018) such as images (Rousseeuw et al., 2018) for which the global ICS could be extended. Further work on the automatic selection of invariant components using a multiple test approach, instead of the scree plot, is also of interest.

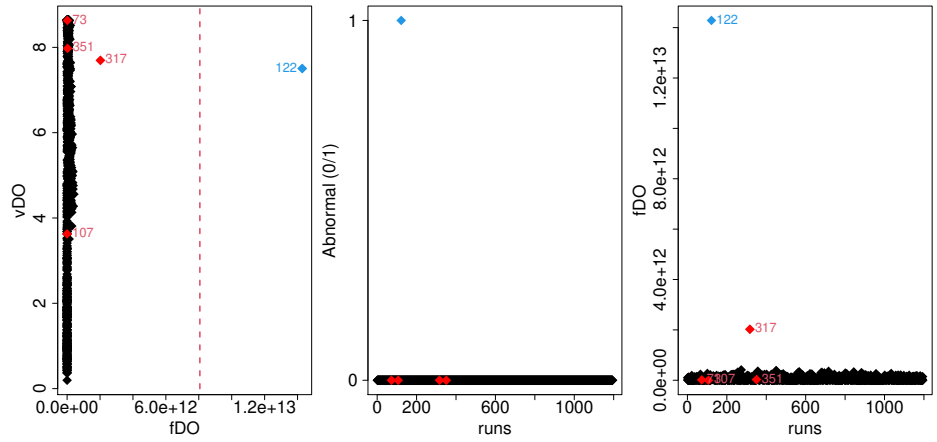

Figure 4.3-22 – **Semiconductor data set** - Directional outlyingness method, Left: FOM using the cutoff quantile of order 0.99999995, Center: outlier flag (1 for abnormal and 0 for normal flights), Right: fDO.

Conclusion

In this part, we apply the invariant component selection method to detect anomalies in multivariate functional data sets for predictive maintenance and quality control.

In Chapter 4, we propose two generalizations of ICS to the functional framework. The first one is point-wise ICS where we apply ordinary ICS at each time point. The second one is global ICS where we apply the ordinary ICS on the projected curves. These approaches can be used in a complementary manner. We show that for the aeronautic data set we were able to detect an abnormal behavior of the generator 4 flights before its.

A perspective of functional ICS on predictive maintenance is to develop an online ICS procedure that will update the eigenvalues and eigenvectors when a new observation is added without having to restart the ICS algorithm. This procedure would allow to track an abnormal behavior of the generator in real time and to raise a warning for maintenance actions. Several applications on online machine learning exist, the work of Feng et al. (2013) on online PCA is an interesting line of future research. We recommend to use on top of online ICS, the oil sample procedure already defined by Airbus which remains essential to optimize the procedure and fine-tune the warning conditions before the in service deployment of the ICS method. The big challenge to deploy this procedure is to automate the components selection step. One possibility already mentioned is to use the asymptotic test developed by Nordhausen et al. (2017) if we have $n > p$. Another possibility is to use algorithms that detect "elbow" or "knee" of a curve (see the works of Satopaa et al. (2011) and Antunes et al. (2018)) as an automatic way to select the components in the scree plot.

Conclusion et perspectives

Durant cette thèse, nous avons utilisé des approches statistiques pour traiter des problématiques liées à la compréhension du comportement des systèmes électriques aéronautiques. Un premier travail concerne l'estimation de la consommation électrique maximale qu'un système électrique doit fournir, dans le but de revoir le dimensionnement des générateurs et certaines limitations. Nous avons utilisé la théorie de valeurs extrêmes pour calculer des quantiles associés à des probabilités très faibles de l'ordre de 10^{-9} (10^{-7} par heure de vol). Les probabilités ont été converties en probabilité par heure de vol pour se conformer aux procédures de sécurité d'Airbus. Nous avons pu estimer des consommations extrêmes et construire des intervalles de confiance en se basant sur les consommations électriques de 18 avions en service. Nous avons comparé les consommations extrêmes aux valeurs théoriques calculées par les ingénieurs et avons constaté qu'elles sont conservatrices dans le régime nominal permanent. La théorie des valeurs extrêmes nous a aussi permis d'estimer des points terminaux et de démontrer que la consommation électrique maximale ne dépend pas de la façon dont l'avion est utilisé. Nous proposons d'étendre cette application sur une plus grande quantité d'avions avec différents types de configurations pour valider la surestimation de l'ELA.

La motivation du deuxième travail réalisé consiste à prédire le comportement du générateur électrique sous des conditions extrêmes (e.g. très haute ou très basse altitude, température extérieure extrême, etc.) que nous ne pouvons pas réaliser physiquement. Même si nous n'avons pas pu atteindre notre objectif initial par manque de temps, nous avons utilisé une approche de représentation fonctionnelle des données pour prédire la température de l'huile à un temps donné en fonction de l'historique des variables explicatives. Nous avons étudié les modèles de machine learning les plus utilisés et avons appliqué une régularisation de type dropout sur nos modèles pour réduire leur variabilité. Nous avons utilisé cette approche pour développer un double digital qui détecte des anomalies dans le comportement du générateur. Le double digital n'est en fait qu'une procédure de machine learning entraînée avec des vols sans anomalies puis testée sur des vols qui contiennent une panne du générateur. Cette approche permet de générer une grande erreur de prédiction pour les vols où le générateur se comporte de façon anormale et de lancer une alerte pour effectuer un contrôle de maintenance.

Enfin, nous avons développé, dans un troisième travail, un modèle de maintenance prédictive basé sur la méthode d'Invariant Coordinate Selection (ICS). Cette méthode présente des propriétés intéressantes pour la détection d'observations atypiques en multivarié dans un contexte non-supervisé. Nous avons adapté les travaux d'Archimbaud et al. (2018) aux données fonctionnelles en utilisant deux approches différentes. La première, appelée point-wise ICS, consiste à faire un ICS standard à chaque point temporel. La deuxième, appelée global ICS, consiste à faire une projection des variables fonctionnelles sur une base orthonormée, puis à appliquer une seule fois l'ICS standard sur une troncature de cette projection. En utilisant cette dernière approche, nous avons pu anticiper les pannes des générateurs électriques. En perspective, nous recommandons d'augmenter le nombre de pannes à tester pour valider la méthode. Il serait aussi souhaitable

d'adapter ICS en ligne pour pouvoir détecter des comportements anormaux en temps réel.

Tout au long de cette thèse, nous avons répondu à des problématiques d'ingénierie très différentes en utilisant des approches statistiques récentes. Nous avons constaté que les ingénieurs sont de plus en plus favorables à la mise en place de solutions de type intelligence artificielle plutôt que des solutions de type reconception (redesign) ou simulation, qui sont souvent plus coûteuses et consommatrices de temps. Grâce à la valeur ajoutée des résultats obtenus dans cette thèse, nous espérons avoir contribué à cette évolution de la vision des designers des systèmes électriques chez Airbus.

Appendices

Oil temperature prediction of aeronautical electrical generator

4.5 Link between dropout regularizer and Tikhonov regularizer

Proof of Equation (2.3.2)**:**

$$\widehat{L}_{\text{drop}}(f_\omega) = \widehat{L}(f_\omega) + \frac{p}{1-p}\|\Gamma \mathrm{w}\|^2 \quad \text{where} \quad \Gamma^2 = \operatorname{ddiag}\left(\mathrm{x}_k \mathrm{x}_k^\top\right).$$

The objective function using the dropout regularization

$$\widehat{L}_{\text{drop}}(f_\omega) = \mathbb{E}_{\mathrm{b}}\left[\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \mathrm{w}^\top \operatorname{diag}(\mathrm{b})\mathrm{x}_k - w_0\right)^2\right],$$

can be developed as follow

$$\begin{aligned}
\widehat{L}_{\text{drop}}(f_\omega) =& \mathbb{E}_{\mathrm{b}}\left[\frac{1}{n}\sum_{k=1}^{n}\left(y_k - \mathrm{w}^\top \mathrm{x}_k - w_0 - \left(\mathrm{w}^\top \operatorname{diag}(\mathrm{b})\mathrm{x}_k - \mathrm{w}^\top \mathrm{x}_k\right)\right)^2\right] \\
=& \frac{1}{n}\sum_{k=1}^{n}\left(y_k - \mathrm{w}^\top \mathrm{x}_k - w_0\right)^2 + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{\mathrm{b}}\left[\left(\mathrm{w}^\top \operatorname{diag}(\mathrm{b})\mathrm{x}_k - \mathrm{w}^\top \mathrm{x}_k\right)^2\right] - \\
& \frac{2}{n}\sum_{k=1}^{n}\left(y_k - \mathrm{w}^\top \mathrm{x}_k - w_0\right)^\top \left(\mathrm{w}^\top \mathbb{E}_{\mathrm{b}}\left[\operatorname{diag}(\mathrm{b})\right]\mathrm{x}_k - \mathrm{w}^\top \mathrm{x}_k\right) \\
=& \widehat{L}(f_\omega) + \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{\mathrm{b}}\left[\left(\mathrm{x}_k^\top \operatorname{diag}(\mathrm{b})\mathrm{w} - \mathrm{x}_k^\top \mathrm{w}\right)\left(\mathrm{w}^\top \operatorname{diag}(\mathrm{b})\mathrm{x}_k - \mathrm{w}^\top \mathrm{x}_k\right)\right] \\
=& \widehat{L}(f_\omega) + \frac{1}{n}\sum_{k=1}^{n}\mathrm{x}_k^\top \mathbb{E}_{\mathrm{b}}\left[\left(\operatorname{diag}(\mathrm{b}) - I\right)\mathrm{w}\mathrm{w}^\top\left(\operatorname{diag}(\mathrm{b}) - I\right)\right]\mathrm{x}_k,
\end{aligned}$$

where $I \in \mathbb{R}^{d\times d}$ stands for the identity matrix.

We recall that $\mathbb{E}_{\mathrm{b}}\left[\operatorname{diag}(\mathrm{b}_i)\right] = 1$ and $\mathbb{E}_{\mathrm{b}}\left[\operatorname{diag}(\mathrm{b}_i)^2\right] = 1/(1-p)$. Thus $\mathbb{E}_{\mathrm{b}}[(\operatorname{diag}(\mathrm{b})_{ii} - I)\mathrm{w}\mathrm{w}^\top(\operatorname{diag}(\mathrm{b})_{jj} -$

$I)] = p/(1-p)w_{ii}^2$ for $i = j$ and 0 otherwise. This implies

$$\begin{aligned}\widehat{L}_{\text{drop}}(f_\omega) =& \widehat{L}(f_\omega) + \frac{p}{(1-p)} \sum_{i=1}^{d} w_i^2 \, \frac{1}{n} \sum_{k=1}^{n} x_{ki}^2 \\ =& \widehat{L}(f_\omega) + \frac{p}{1-p} \|\Gamma \mathrm{w}\|^2,\end{aligned}$$

where $\Gamma^2 = \text{ddiag}\left(\frac{1}{n} \sum_{k=1}^{n} \mathrm{x}_k \mathrm{x}_k^\top\right)$.

Proof of Equation (2.3.4)**:**

$$\widehat{L}_{\text{drop}}(f_\omega) = \widehat{L}(f_\omega) + \frac{p}{1-p} \left\| \Gamma \left(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)} \right) \mathrm{w} \right\|^2$$

where

$$\Gamma^2 \left(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)} \right) = \text{ddiag} \left(\frac{1}{n} \sum_{k=1}^{n} \varphi \left(\mathrm{V}^{(1)} \mathrm{x}_k + \mathrm{v}_0^{(1)} \right) \varphi \left(\mathrm{V}^{(1)} \mathrm{x}_k + \mathrm{v}_0^{(1)} \right)^\top \right).$$

We denote $\Phi_k = \varphi \left(\mathrm{V}^{(1)} \mathrm{x}_k + \mathrm{v}_0^{(1)} \right) \in \mathbb{R}^{d_1}$, the objective function of single layer perceptron using the dropout regularization

$$\begin{aligned}\widehat{L}_{\text{drop}}(f_\omega) &= \mathbb{E}_{\mathrm{b}} \left[\frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \text{diag}(\mathrm{b}) \Phi_k - w_0 \right)^2 \right] \\ &= \mathbb{E}_{\mathrm{b}} \left[\frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \Phi_k - w_0 - \left(\mathrm{w}^\top \text{diag}(\mathrm{b}) \Phi_k - \mathrm{w}^\top \Phi_k \right) \right)^2 \right] \\ &= \frac{1}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \Phi_k - w_0 \right)^2 + \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}_{\mathrm{b}} \left[\left(\mathrm{w}^\top \text{diag}(\mathrm{b}) \Phi_k - \mathrm{w}^\top \Phi_k \right)^2 \right] - \\ &\quad \frac{2}{n} \sum_{k=1}^{n} \left(y_k - \mathrm{w}^\top \Phi_k - w_0 \right)^\top \mathbb{E}_{\mathrm{b}} \left[\left(\mathrm{w}^\top \text{diag}(\mathrm{b}) \Phi_k - \mathrm{w}^\top \Phi_k \right)^2 \right] \\ &= \widehat{L}(f_\omega) + \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}_{\mathrm{b}} \left[\left(\mathrm{w}^\top \text{diag}(\mathrm{b}) \Phi_k - \mathrm{w}^\top \Phi_k \right)^2 \right].\end{aligned}$$

Similarly to the previous proof

$$\begin{aligned}\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}_{\mathrm{b}} \left[\left(\mathrm{w}^\top \text{diag}(\mathrm{b}) \Phi_k - \mathrm{w}^\top \Phi_k \right)^2 \right] &= \frac{1}{n} \sum_{k=1}^{n} \Phi_k^\top \mathbb{E}_{\mathrm{b}} \left[(\text{diag}(\mathrm{b}) - I) \, \mathrm{w}\mathrm{w}^\top \, (\text{diag}(\mathrm{b}) - I) \right] \Phi_k \\ &= \frac{p}{1-p} \sum_{i=1}^{d_1} w_i^2 \frac{1}{n} \sum_{k=1}^{n} \phi_{ki}^2, \text{ where } \phi_{ki} = \varphi \left(\mathrm{V}^{(1)} \mathrm{x}_k + \mathrm{v}_0^{(1)} \right)_i \\ &= \frac{p}{1-p} \left\| \Gamma \left(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)} \right) \mathrm{w} \right\|^2,\end{aligned}$$

where $\Gamma^2\left(\mathrm{V}^{(1)}, \mathrm{v}_0^{(1)}\right) = \mathrm{ddiag}\left(\frac{1}{n}\sum_{k=1}^{n}\Phi_k\Phi_k^\top\right)$.

Abnormal behavior detection of aeronautical electrical generator

4.6 ICS for multivariate functional anomaly detection with applications to predictive maintenance and quality control

4.6.1 Supplementary figures for the Weather data set

Visualization of the outliers on the aligned curves In Figures 4.6-23 to 4.6-25 we give the curves for the three variables, and we highlight the abnormal curves by variable and add the curves number. In Figures 4.6-23 (resp., 4.6-24 and 4.6-25) we coloured in red (resp., blue and green) the abnormal curves observed in the temperature (resp., wind and log precipitation) on the 3 variables.

4.6.2 Supplementary figures for the Aeronautic data set

Visualization of the outliers on the aligned curves In Figure 4.6-26, we plot 100 normal flights in grey and highlight the four abnormal flights that precede the generator loss in red. It can be seen that the four red curves differ from the others at the end of the flight with an engine speed that drops compared to the other curves.

Global ICS application Figure 4.6-27 gives the correlations between the first invariant component and the 5 Fourier coefficients of the 6 initial variables. The first component is clearly only correlated with coefficients associated with the engine speed (X^3).

Point-wise ICS application The point-wise ICS results where we automatically select the number of components by time using the test proposed by Nordhausen et al. (2017) but do not

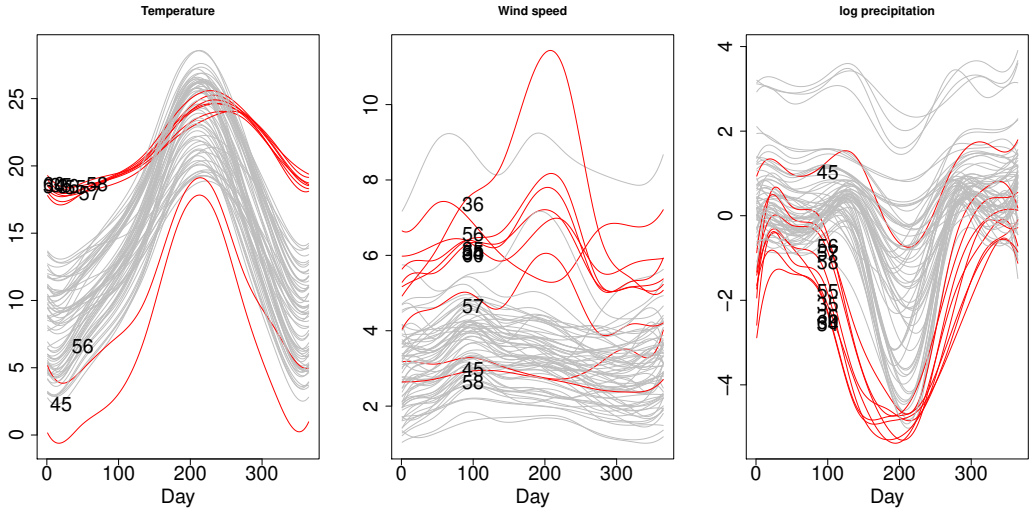

Figure 4.6-23 – **Weather data set** - Suspected outlying curves (34,35,36,55,56,57,58,60,45) flagged in red based on the temperature.

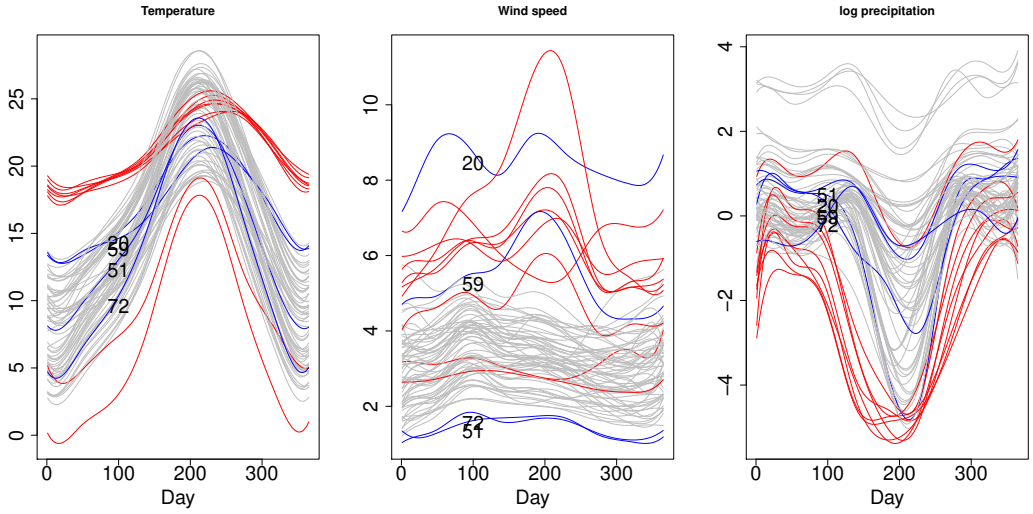

Figure 4.6-24 – **Weather data set** - Additional suspected outlying curves (20,59,51,72) flagged in blue based on the wind speed.

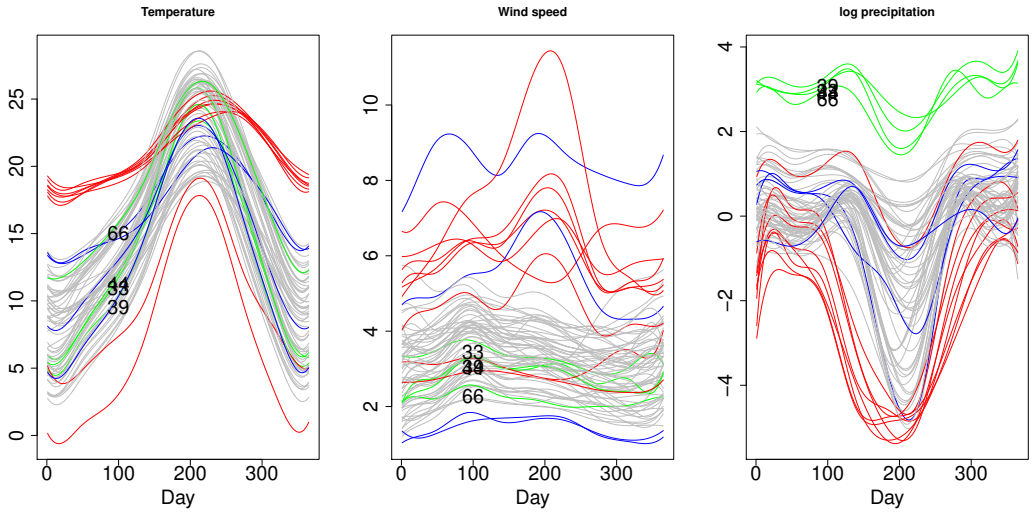

Figure 4.6-25 – **Weather data set** - Additional suspected outlying curves (33,39,44,66) flagged in green based on the log precipitation.

divide the ICS distances by the number of components selected is given in Figure 4.6-28. The 4 flights preceding the generator loss are in red and the additional flights detected as abnormal by point-wise ICS are in blue. Compared to the weighted point-wise ICS we detect more false abnormal flights.

In Figures 4.6-29 and 4.6-30 we give the distance by flight and by time using $k = 1$ at each time point. We note that, the weighted point-wise ICS and the point-wise ICS with $k = 1$ give similar results.

Directional outlyingness method The result of the directional outlyingness method proposed by Rousseeuw et al. (2018), without changing the *distOptions* in the *fom* function is given in Figure 4.6-31. The four flights that precede the generator loss are coloured in red, the abnormal flights detected by the FOM cutoff on the vDO axis are coloured in green and those detected by the fDO axis are coloured in blue.

The abnormal flights detected are very different from the flights detected using global or point-wise ICS. Figure 4.6-32 gives the time points where the weight is equal to zero. In fact, all of the cruise period is removed from the analyisis.

4.6.3 Supplementary figures for the Semiconductor data set

Visualization of the outliers on the aligned curves In Figure 4.6-33, we plot in grey the normal curves and in red the abnormal curves. It is not esay to interpret the outlyingness of the red curves but it seems that there is a mix between shift and amplitude outliers.

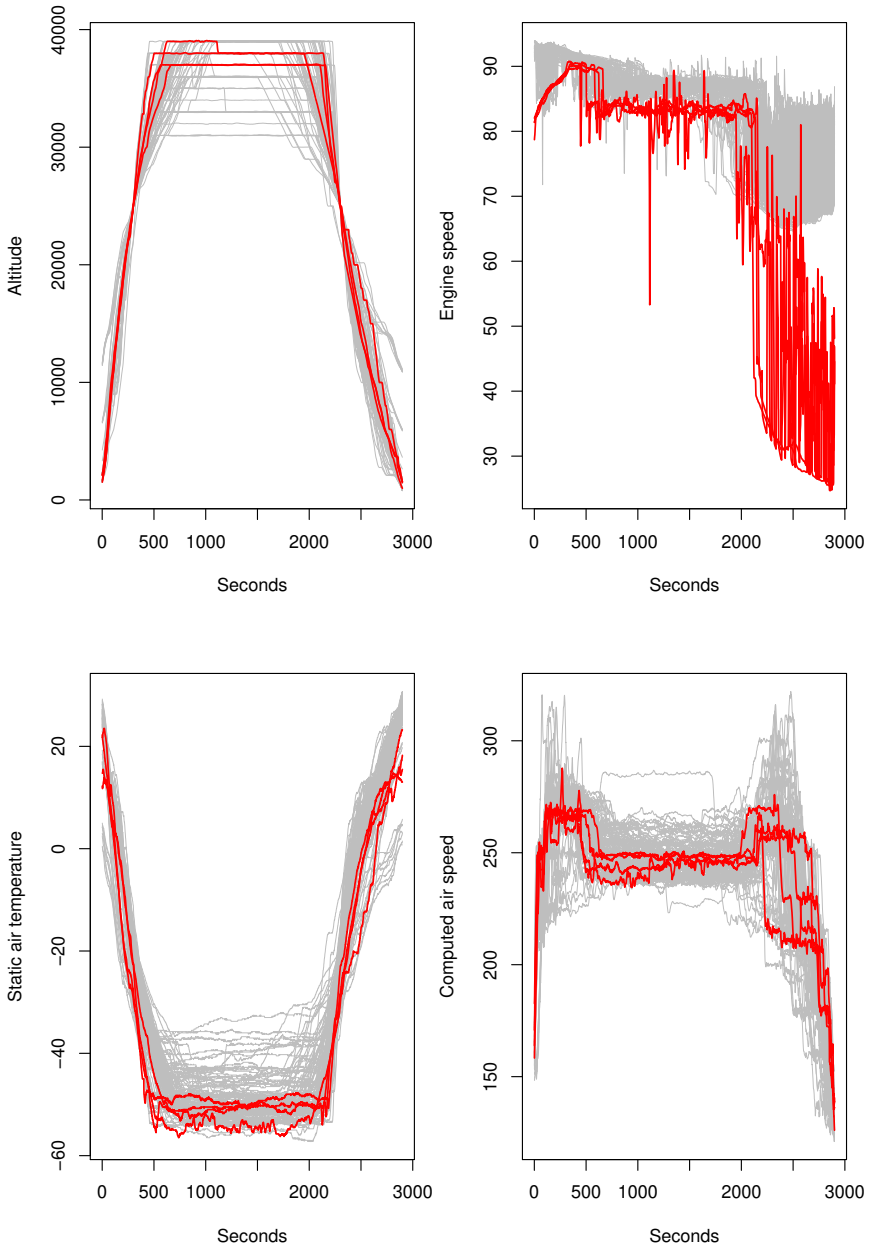

Figure 4.6-26 – **Aeronautical data set** - Observed flights after alignment by feature with $T = 2900$ seconds. In red the flights that precede the generator loss.

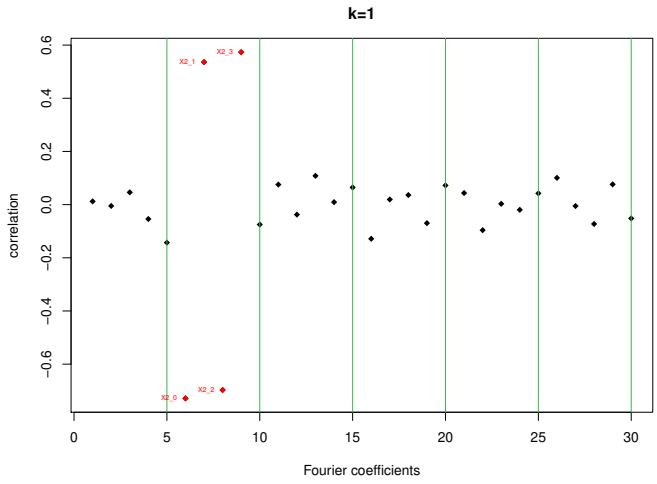

Figure 4.6-27 – **Aeronautical data set** - Global ICS: correlations between the first invariant component and the 5 Fourier coefficients of the 6 initial variables. Red colour corresponds to correlation with absolute value larger than 0.20. The green lines are separators between the features which are ordered from X^1 to X^6.

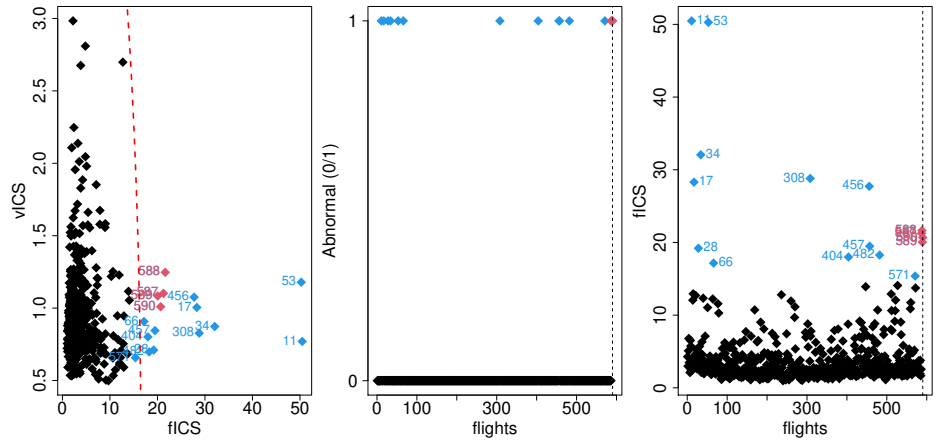

Figure 4.6-28 – **Aeronautical data set** - Point-wise ICS, Left: FOM using the cutoff quantile of order 0.999995, Center: outlier flag (1 for abnormal and 0 for normal flights), Right: fICS.

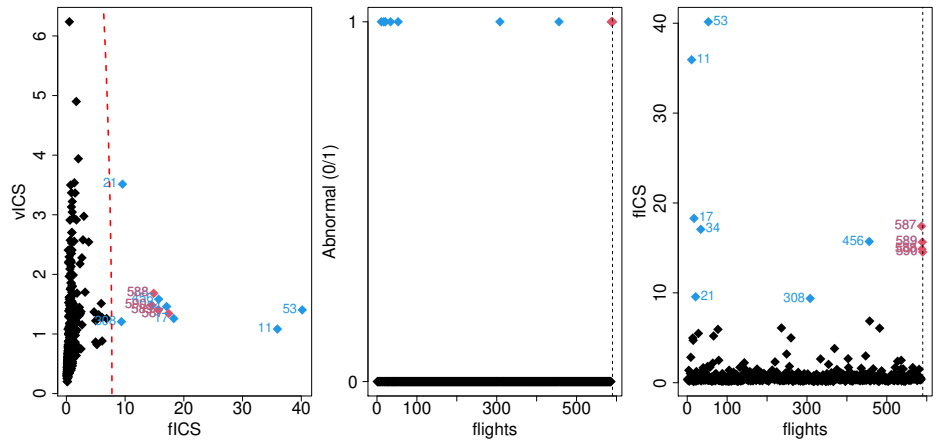

Figure 4.6-29 – **Aeronautical data set** - Point-wise ICS (with $k = 1$), Left: FOM using the cutoff quantile of order 0.99999995, Center: outlier flag (1 for abnormal and 0 for normal flights), Right: fICS.

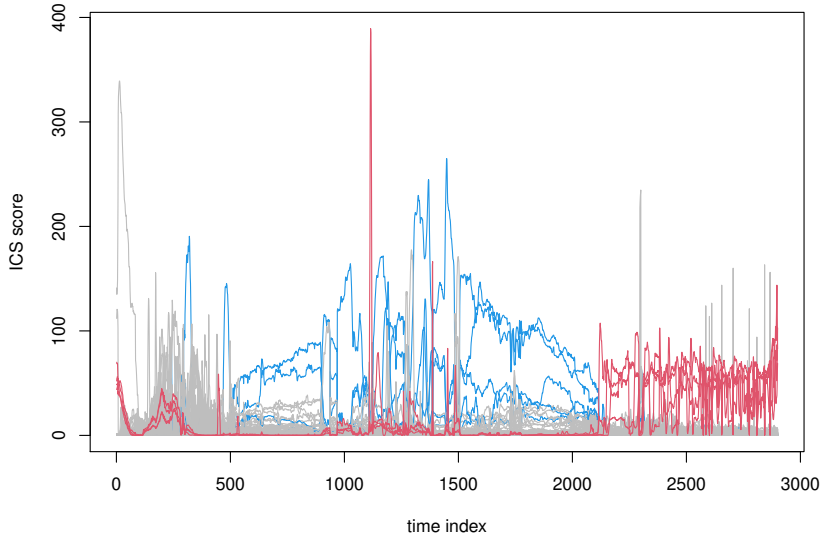

Figure 4.6-30 – **Aeronautical data set** - Point-wise ICS: square ICS distance ($k = 1$). In red the 4 flights that precede failure and in blue the flights detected as outliers besides the red ones.

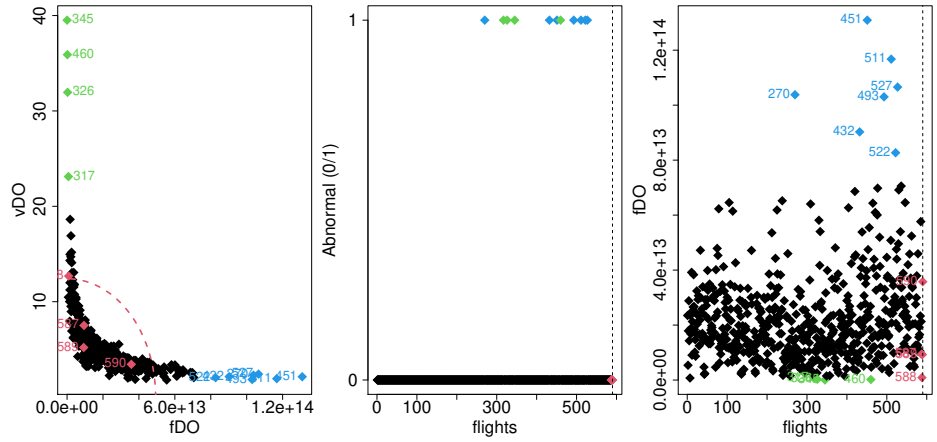

Figure 4.6-31 – **Aeronautical data set** - Directional outlyingness method, Left: FOM using the default cutoff quantile of 0.995. Center: clustering of flights, 1 for abnormal and 0 for normal flights. Right: fDO.

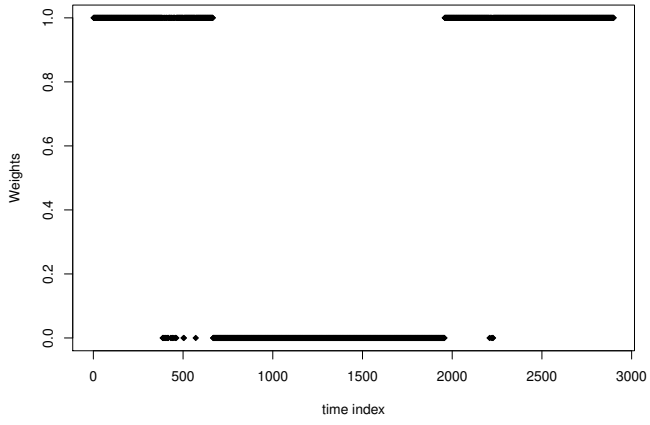

Figure 4.6-32 – **Aeronautical data set** - Directional outlyingness method: time points where the weights are equal to zero (46%).

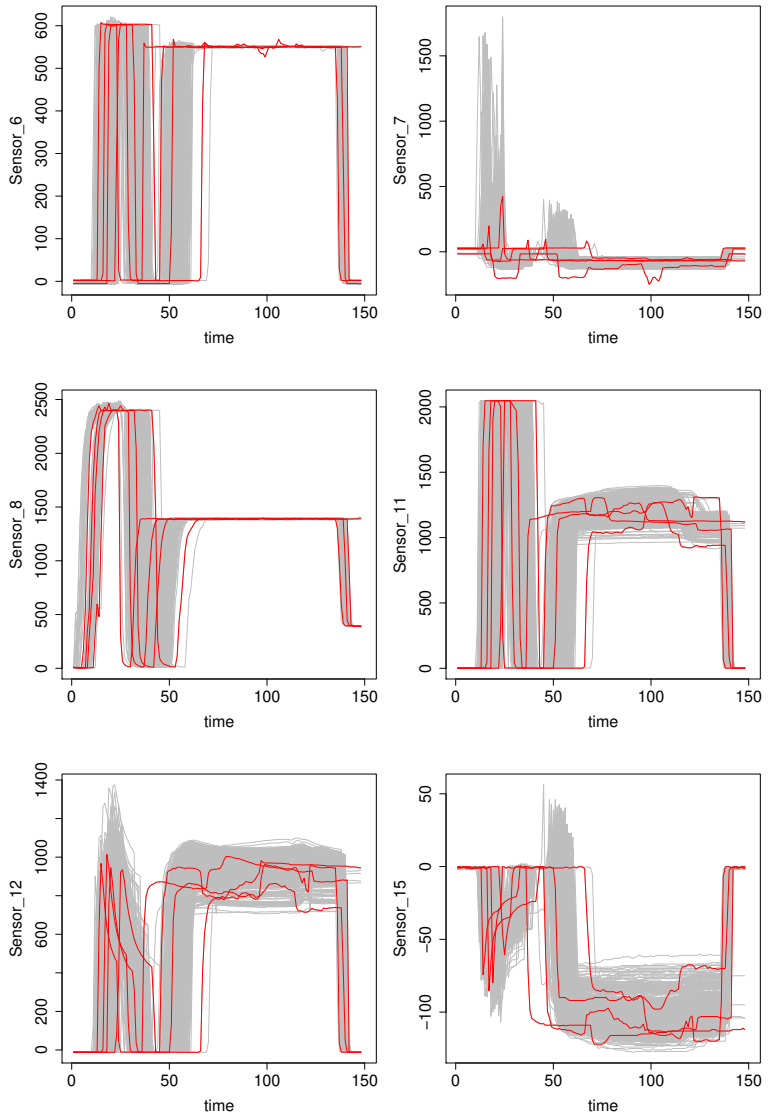

Figure 4.6-33 – **Semiconductor data set** - Observed runs by sensor after alignment. Abnormal runs are coloured in red.

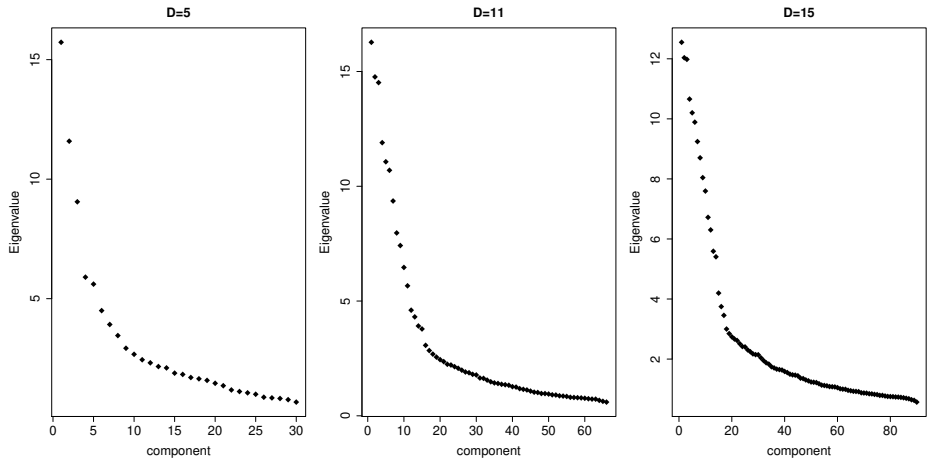

Figure 4.6-34 – **Semiconductor data set** - Global ICS: scree plot with $D = 5$, 11 and 15 coefficients of the Fourier basis.

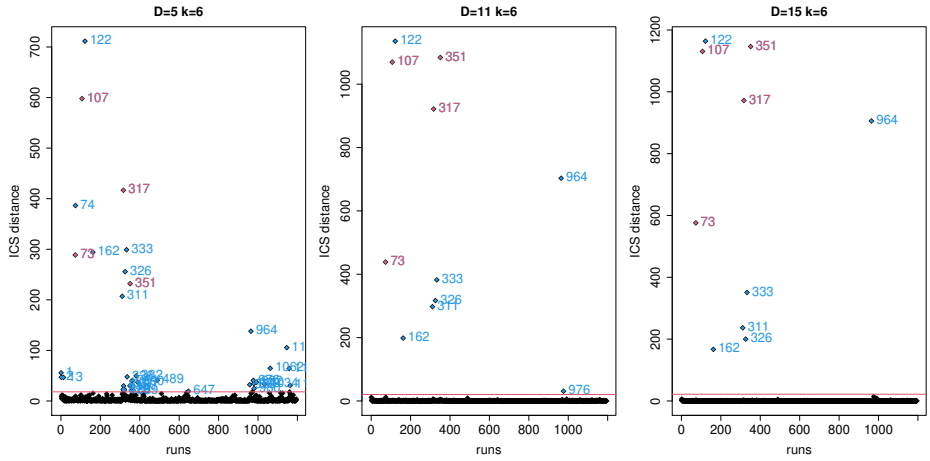

Figure 4.6-35 – **Semiconductor data set** - Global ICS: square ICS distance with $D = 5$, 11 and 15 coefficients of the Fourier basis and $k = 6$.

Global ICS In Figures 4.6-34 and 4.6-35 respectively we give the scree plot and the ICS distances by run using the first 6 ICS components for each $D = 5, 11$ and 15. The results for $D = 11$ and 15 are similar, but for $D = 5$, we detect a large number of false positive runs. This finding can be explained by a loss of information when we select a low truncation.

Directional outlyingness method We observe in Figure 4.6-36 the issue we already mentioned with the *fom* function and the 50% of the weights that are equal to zero.

Finally, Figure 4.6-37 gives the results of the directional outlyingness method when changing the *distOptions* but they do not differ a lot compared with the results with the default options.

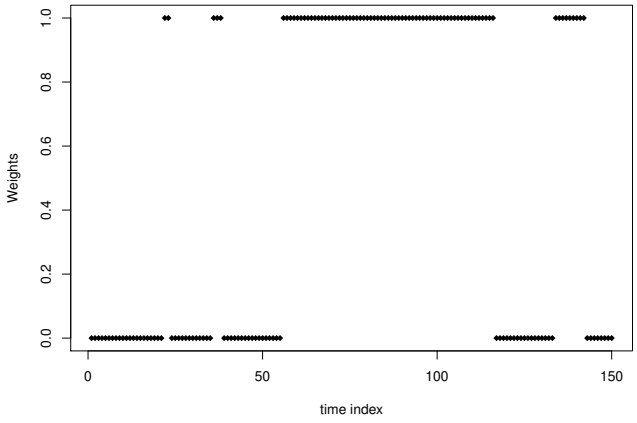

Figure 4.6-36 – **Semiconductor data set** - Directional outlyingness method: time points where the weights are equal to zero (50%).

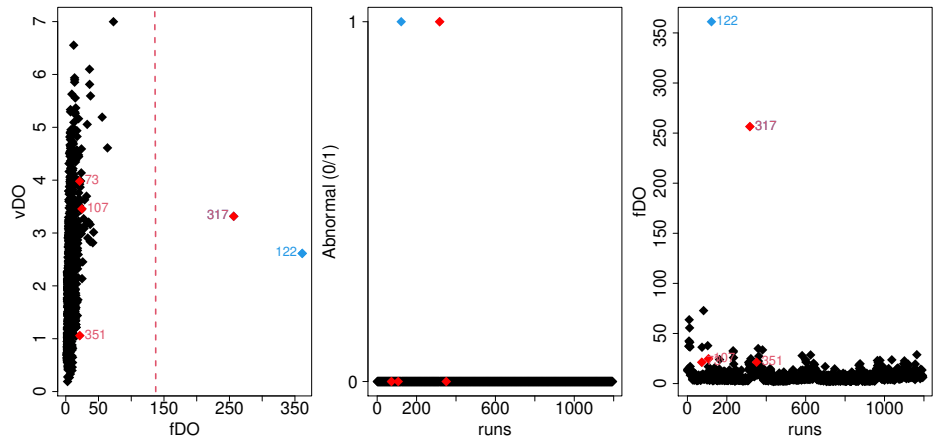

Figure 4.6-37 – **Semiconductor data set** - Directional outlyingness method with distOptions = list(rmZeroes = TRUE,maxRatio = 3) in the fOutl function. Left: FOM, Center: outlier flag (1 for abnormal and 0 for normal flights), Right: fDO.

Bibliography

Martín Abadi and *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015.

Panagiotis Aivaliotis, Konstantinos Georgoulias, and Kosmas Alexopoulos. Using digital twin for maintenance applications in manufacturing: State of the art and gap analysis. In *2019 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, pages 1–5. IEEE, 2019.

M. Antunes, D. Gomes, and R. L. Aguiar. Knee/elbow estimation based on first derivative threshold. In *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 237–240, 2018. doi: 10.1109/BigDataService.2018.00042.

A. Archimbaud, K. Nordhausen, and A. Ruiz-Gazen. ICS for multivariate outlier detection with application to quality control. *Computational Statistics and Data Analysis*, 128:184–199, 2018.

Aurore Archimbaud. Détection non-supervisée d'observations atypiques en contrôle de qualité: un survol. *Journal de la Société Française de Statistique*, 159(3):1–39, 2018a.

Aurore Archimbaud. *Méthodes statistiques de détection d'observations atypiques pour des données en grande dimension*. PhD thesis, Toulouse 1, 2018b.

Aurore Archimbaud, Joris May, Klaus Nordhausen, and Anne Ruiz-Gazen. *ICSShiny: ICS via a Shiny Application*, 2018a. URL `https://CRAN.R-project.org/package=ICSShiny`. R package version 0.5.

Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen. *ICSOutlier: Outlier Detection Using Invariant Coordinate Selection*, 2018b. URL `https://CRAN.R-project.org/package=ICSOutlier`. R package version 0.3-0.

Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen. Unsupervized outlier detection with icsoutlier. *The R Journal*, 10(1):234–250, 2018c.

Raman Arora, Peter Bartlett, Poorya Mianjy, and Nathan Srebro. Dropout: Explicit Forms and Capacity Control, 2020.

U Raghu Babu and B Kondraivendhan. Application of statistics to the analysis of corrosion data for rebar in metakaolin concrete. In *International Conference on Emerging Trends in Engineering (ICETE)*, pages 162–169. Springer, 2020.

Francis Bach and Eric Moulines. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Gray Bachelor, Eugenio Brusa, Davide Ferretto, and Andreas Mitschke. Model-based design of complex aeronautical systems through digital twin and thread concepts. *IEEE Systems Journal*, 2019.

Clémentine Barreyre, Loic Boussouf, Bertrand Cabon, Béatrice Laurent, and Jean-Michel Loubes. Statistical methods for outlier detection in space telemetries. In *Space Operations: Inspiring Humankind's Future*, pages 513–547. Springer, 2019.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Christoph Norbert Bergmeir and José Manuel Benítez Sánchez. Neural networks in r using the stuttgart neural network simulator: Rsnns. *Journal of Statistical Software*, 46, 2012.

Ufuk Beyaztas and Zaher Mundher Yaseen. Drought interval simulation using functional data analysis. *Journal of Hydrology*, 579:124141, 2019.

Chris Bishop. Training with Noise is Equivalent to Tikhonov Regularization. *Neural Computation*, 7(1):108–116, 1995.

F. Boulfani, X. Gendre, A. Ruiz-Gazen, and M. Salvignol. Anomaly detection for aircraft electrical generator using machine learning in a functional data framework. In *2020 Global Congress on Electrical Engineering (GC-ElecEng)*, pages 27–32, 2020. doi: 10.23919/GC-ElecEng48342.2020.9285984.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. 1984.

Enrique Castillo, Ali S Hadi, Narayanaswamy Balakrishnan, and José-Mariá Sarabia. *Extreme value and related models with applications in engineering and science*. Wiley Hoboken, NJ, 2005.

Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

Wenlin Dai and Marc G Genton. Multivariate functional data visualization and outlier detection. *Journal of Computational and Graphical Statistics*, 27(4):923–934, 2018.

Wenlin Dai and Marc G Genton. Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, 131:50–65, 2019.

Wenlin Dai, Tomáš Mrkvička, Ying Sun, and Marc G Genton. Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics & Data Analysis*, page 106960, 2020.

Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2006.

Jesson J Einmahl, John HJ Einmahl, and Laurens de Haan. Limits to human life span through extreme value theory. *Journal of the American Statistical Association*, 114(527):1075–1080, 2019.

S. El Adlouni, T. B. M. J. Ouarda, X. Zhang, R. Roy, and B. Bobée. Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research*, 43(3), 2007.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

Bircan Erbas, Rob J Hyndman, and Dorota M Gertig. Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine*, 26(2):458–470, 2007.

Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Online pca for contaminated data. *Advances in Neural Information Processing Systems*, 26:764–772, 2013.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

David Ganger, Junshan Zhang, and Vijay Vittal. Statistical characterization of wind power ramps via extreme value analysis. *IEEE Transactions on Power Systems*, 29(6):3118–3119, 2014.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

Eric Gilleland, Richard W Katz, et al. extremes 2.0: an extreme value analysis package in r. *Journal of Statistical Software*, 72(8):1–39, 2016.

Manfred Gilli et al. An application of extreme value theory for measuring financial risk. *Computational Economics*, 27(2-3):207–228, 2006.

Xavier Giraud. *Méthodes et outils pour la conception optimale des réseaux de distribution d'électricité dans les aéronefs*. PhD thesis, Toulouse, INSA, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Margaret E Gruen, Marcela Alfaro-Córdoba, Andrea E Thomson, Alicia C Worth, Ana-Maria Staicu, and B Duncan X Lascelles. The use of functional data analysis to evaluate activity in a spontaneous model of degenerative joint disease associated pain in cats. *PloS one*, 12(1): e0169576, 2017.

Alexander Hagg, Maximilian Mensing, and Alexander Asteroth. Evolving Parsimonious Networks by Mixing Activation Functions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO'17, pages 425–432, 2017.

Clara Happ and Sonja Greven. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113(522):649–659, 2018.

Trevor Harris, J Derek Tucker, Bo Li, and Lyndsay Shand. Elastic depths for detecting shape anomalies in functional data. *Technometrics*, just-accepted:1–25, 2020.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning, second edition*. Springer, 2009.

Christina Heinze, Brian McWilliams, Nicolai Meinshausen, and Gabriel Krummenacher. LOCO: Distributing Ridge Regression with Random Projections, 2014.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Mia Hubert, Peter J Rousseeuw, and Pieter Segaert. Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2):177–202, 2015.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Gabriel Jarry, Daniel Delahaye, Florence Nicol, and Eric Feron. Aircraft atypical approach detection using functional principal component analysis. *Journal of Air Transport Management*, 84: 101787, 2020.

Joarder Kamruzzaman and Syed Mahfuzul Aziz. A note on activation function in multilayer feedforward learning. In *Proceedings of the 2002 International Joint Conference on Neural Networks*, volume 1 of *IJCNN'02*, pages 519–523, 2002.

Sonja Kuhnt and André Rehage. An angle-based multivariate functional pseudo-depth for shape outlier detection. *Journal of Multivariate Analysis*, 146:325–340, 2016.

Jordan Larson and Demoz Gebre-Egziabher. Conservatism assessment of extreme value theory overbounds. *IEEE Transactions on Aerospace and Electronic Systems*, 53(3):1295–1307, 2017.

Algirdas Laukaitis. Functional data analysis for cash flow and transactions intensity continuous-time prediction using hilbert-valued autoregressive processes. *European Journal of Operational Research*, 185(3):1607–1614, 2008.

Clément Lejeune, Josiane Mothe, Adil Soubki, and Olivier Teste. Shape-based outlier detection in multivariate functional data. *Knowledge-Based Systems*, page 105960, 2020.

Bing Li, Germain Van Bever, Hannu Oja, Radka Sabolová, and Frank Critchley. Functional independent component analysis: an extension of the fourth-order blind identification, 2019.

Lishuai Li, Santanu Das, R John Hansman, Rafael Palacios, and Ashok N Srivastava. Analysis of flight data using clustering techniques for detecting abnormal operations. *Journal of Aerospace Information Systems*, 12(9):587–598, 2015.

Lishuai Li, R John Hansman, Rafael Palacios, and Roy Welsch. Anomaly detection via a gaussian mixture model for flight operation and safety monitoring. *Transportation Research Part C: Emerging Technologies*, 64:45–57, 2016.

Jingxiang Liu, Junghui Chen, and Dan Wang. Wavelet functional principal component analysis for batch process monitoring. *Chemometrics and Intelligent Laboratory Systems*, 196:103897, 2020.

Milad Memarzadeh, Bryan Matthews, and Ilya Avrekh. Unsupervised anomaly detection in flight data using convolutional variational auto-encoder. *Aerospace*, 7(8):115, 2020.

H. N. Mhaskar and C. A. Micchelli. How to Choose an Activation Function. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pages 319–326, 1993.

Poorya Mianjy, Raman Arora, and Rene Vidal. On the Implicit Bias of Dropout. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 3540–3548, 2018.

Laura Millán-Roures, Irene Epifanio, and Vicente Martínez. Detection of anomalies in water networks by functional data analysis. *Mathematical Problems in Engineering*, 2018, 2018.

Julia E Morrison and James A Smith. Stochastic modeling of flood peaks using the generalized extreme value distribution. *Water Resources Research*, 38(12):41–1, 2002.

Stanislav Nagy, Irène Gijbels, and Daniel Hlubinka. Depth-based recognition of shape outlying functions. *Journal of Computational and Graphical Statistics*, 26(4):883–893, 2017.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer, 2013.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-Based Capacity Control in Neural Networks. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *PMLR*, pages 1376–1401, 2015.

Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

Klaus Nordhausen and David E Tyler. A cautionary note on robust covariance plug-in methods. *Biometrika*, 102(3):573–588, 2015.

Klaus Nordhausen, Hannu Oja, David E Tyler, and Joni Virta. Asymptotic and bootstrap tests for the dimension of the non-gaussian subspace. *IEEE Signal Processing Letters*, 24(6):887–891, 2017.

Robert T Olszewski. Generalized feature extraction for structural pattern recognition in time-series data. Technical report, Carnegie-Mellon Univ. Pittsburgh PA School of Computer Science, 2001.

P Papachatzakis, N Papakostas, and G Chryssolouris. Condition based operational risk assessment an innovative approach to improve fleet and aircraft operability: Maintenance planning. In *1st European Air and Space Conference, Berlin, Germany*, pages 121–126, 2007.

James Pickands III et al. Statistical inference using extreme order statistics. *the Annals of Statistics*, 3(1):119–131, 1975.

Boris Polyak and Anatoli Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Virgilio Quintana, Louis Rivest, Robert Pellerin, Frédérick Venne, and Fawzi Kheddouci. Will model-based definition replace engineering drawings throughout the product lifecycle? a global perspective from aerospace industry. *Computers in industry*, 61(5):497–508, 2010.

U. Radojicic and K. Nordhausen. Non-gaussian component analysis: Testing the dimension of the signal subspace. In M. Maciak, M. Pesta, and M. Schindler, editors, *Analytical Methods in Statistics. AMISTAT 2019*, pages 101–123. Springer, Cham, 2020.

James Ramsay and BW Silverman. *Functional Data Analysis*. Springer Science & Business Media, 2005.

Sarah J Ratcliffe, Leo R Leader, and Gillian Z Heller. Functional data analysis with application to periodically stimulated foetal heart rate data. i: Functional regression. *Statistics in Medicine*, 21(8):1103–1114, 2002.

Geoffroy Roblot. *Méthodologie de pré-dimensionnement de la puissance électrique des générateurs d'un réseau embarqué à partir d'analyses statistiques des consommateurs*. PhD thesis, Nantes University, 2012.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Peter J Rousseeuw, Jakob Raymaekers, and Mia Hubert. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, 27(2):345–359, 2018.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

David Ruppert. Efficient Estimations from a Slowly Convergent Robbins-Monro Process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011.

Ravinka Seresinhe and Craig Lawson. Electrical load-sizing methodology to aid conceptual and preliminary design of large commercial aircraft. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 229(3):445–466, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Guillaume Staerman, Pavlo Mozharovskyi, Stephan Clémençon, and Florence d'Alché Buc. Functional isolation forest. *arXiv preprint arXiv:1904.04573*, 2019.

Jamaludin Suhaila, Abdul Aziz Jemain, Muhammad Fauzee Hamdan, and Wan Zawiah Wan Zin. Comparing rainfall patterns between regions in peninsular malaysia via a functional data analysis technique. *Journal of hydrology*, 411(3-4):197–206, 2011.

Xu Sun, Jian Shi, Shaoping Wang, and Chao Zhang. Design of load spectrum for hydraulic pumps based on extreme value theory. In *2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1321–1326. IEEE, 2017.

Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6): 1947–1958, 2003.

Javier Martínez Torres, Jorge Pastor Pérez, Joaquín Sancho Val, Aonghus McNabola, Miguel Martínez Comesaña, and John Gallagher. A functional data analysis approach for the detection of air pollution episodes and outliers: A case study in dublin, ireland. *Mathematics*, 8(2):225, 2020.

J. Derek Tucker. *fdasrvf: Elastic Functional Data Analysis*, 2020. URL `https://CRAN.R-project.org/package=fdasrvf`. R package version 1.9.4.

J Derek Tucker, Wei Wu, and Anuj Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66, 2013.

Eric Tuegel. The airframe digital twin: some challenges to realization. In *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA*, page 1812, 2012.

David E. Tyler, Frank Critchley, Lutz Dümbgen, and Hannu Oja. Invariant coordinate selection. *Journal of the Royal Statistical Society: Series B*, 71(3):549–592, 2009.

Amrit Shankar Verma, Zhen Gao, Zhiyu Jiang, Zhengru Ren, and Nils Petter Vedvik. Structural safety assessment of marine operations from a long-term perspective: A case study of offshore wind turbine blade installation. In *ASME 2019 38th International Conference on Ocean, Offshore and Arctic Engineering*. American Society of Mechanical Engineers Digital Collection, 2019.

Joni Virta, Bing Li, Klaus Nordhausen, and Hannu Oja. Independent component analysis for multivariate functional data. *Journal of Multivariate Analysis*, 176:104568, 2020.

Stefan Wager, Sida Wang, and Percy Liang. Dropout training as adaptive regularization. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Sida Wang and Christopher Manning. Fast dropout training. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Per Westerlund and Wadih Naim. Extreme value analysis of power system data. In *ITISE 2019-International Conference on Time Series and Forecasting, 25-27 September 2019 Granada (Spain)*, volume 1, pages 322–327, 2019.

Résumé

La caractérisation des systèmes électriques est une tâche essentielle dans la conception aéronautique. Elle consiste notamment à dimensionner les composants des systèmes, définir les exigences à respecter par les charges électriques, définir les intervalles de maintenance et identifier les causes racines des pannes sur avions. Aujourd'hui, les calculs sont basés sur la théorie du génie électrique ou des modèles physiques simulés. L'objectif de cette thèse est d'utiliser une approche statistique basée sur les données observées durant les vols et des modèles d'apprentissage automatique pour caractériser le comportement du système électrique aéronautique. La première partie de cette thèse traite de l'estimation de la consommation électrique maximale que fournit un système électrique, dans le but d'optimiser le dimensionnement des générateurs et de mieux connaître les marges réelles. La théorie des valeurs extrêmes a été utilisée pour estimer des quantiles qui sont comparés aux valeurs théoriques calculées par les ingénieurs. Dans la deuxième partie, différents modèles régularisés sont considérés pour prédire la température de l'huile du générateur électrique dans un contexte de données fonctionnelles. Ces modèles, appliqués à des données de vols sans anomalie, permettent notamment de détecter des comportements anormaux du générateur associés à de grandes erreurs de prédiction. Enfin, dans la dernière partie, un modèle de maintenance prédictive est proposé afin de détecter des anomalies dans le fonctionnement du générateur électrique pour anticiper les pannes. Le modèle proposé utilise des variantes de la méthode "Invariant Coordinate Selection" pour des données fonctionnelles.

Mots-clés : théorie des valeurs extrêmes ; données fonctionnelles multivariées ; apprentissage automatique ; détection d'anomalies ; maintenance prédictive ; analyse de données de vol.

Abstract

The characterization of electrical systems is an essential task in aeronautic conception. It consists in particular of sizing the electrical components, defining maintenance frequency and finding the root cause of aircraft failures. Nowadays, the computations are made using electrical engineering theory and simulated physical models. The aim of this thesis is to use statistical approaches based on flight data and machine learning models to characterize the behavior of aeronautic electrical systems. In the first part, we estimate the maximal electrical consumption that the generator should deliver to optimize the generator size and to better understand its real margin. Using the extreme value theory we estimate quantiles that we compare to the theoretical values computed by the electrical engineers. In the second part, we compare different regularized procedures to predict the oil temperature of a generator in a functional data framework. In particular, these models applied to anomaly-free flight data make it possible to detect abnormal generator behaviour associated with large prediction errors. Finally, in the last part, we develop a predictive maintenance model that detects the abnormal behavior of a generator to anticipate failures. This model is based on variants of "Invariant Coordinate Selection" adapted to functional data.

Keywords: extreme value theory ; multivariate functional data ; machine learning ; anomaly detection ; predictive maintenance ; flight data analysis.